Nonparametric Statistics with Shape Constraints and Censored Data

Kim Hendrickx

Promoter: Prof. dr. Anneleen Verhasselt

Co-promoter: Prof. dr. em. Paul Janssen

Co-promoter: Prof. dr. em. Piet Groeneboom

Acknowledgements

During my PhD, I have spent a lot of time on writing and answering, at every possible hour of the day, e-mails, “PS-mails" and even “PS2,PS3,...-mails” to my co-supervisor Piet Groeneboom. Piet, our conversations were often very fruitful leading to an overload of nice ideas and interesting research topics. I benefited a lot from your expertise and know-how. You were always willing to help me and introduced me to your broad network of statisticians. Not only did you take care of my academic growth, you also made sure that I increased my knowledge of music and literature and our monthly meetings at Delft even improved my driving skills. I am very grateful for all the opportunities that you offered me.

I also want to express my gratitude to my other supervisors Anneleen Verhasselt and Paul Janssen who made it possible for me to start my PhD project. Thank you Anneleen for offering me my PhD position and for giving me the opportunity to collaborate with Piet. Paul, I will never forget the patience you had each time when I had smaller or bigger problems that needed solutions. You always succeeded in finding constructive solutions.

My next sincere thanks go to all the colleagues at Censtat. It was always a pleasure to meet you for lunch and coffee breaks, in the hallways, ... I enjoyed working in this nice Censtat environment. Thanks also to all JOSS members. I am really proud that I was part of this nice group of enthusiastic colleagues who even gave me the opportunity to be president of JOSS for the last two years of my PhD. Julie, Zoë, Andrea, Mohammed, Tapiwa, Thao, Tran, Alvaro,.. I really had a lot of fun during our nice (after-)work hours. Thank you Sereina for listening, supporting and motivating me during my PhD. I still remember the first time we met and how a beautiful friendship started. You are always there for me with great advise and nice ideas. You will always be more than welcome to stay with us!

Mama en papa, bedankt dat jullie me de kans gegeven hebben om te studeren. Zonder jullie was ik hier nooit aan begonnen. De voorbije jaren zijn niet altijd gemakkelijk geweest maar als er één constante was, is het toch wel dat jullie altijd voor me klaar stonden en ik ben ervan overtuigd dat dit in de toekomst niet anders zal zijn. Mama, "Das love he" ;-). Robin, je bent een geweldige broer en in tijden van nood kon ik altijd rekenen op een hartvermarmende knuffel van jou! Ik wil ook de rest van mijn familie bedanken, nonkel Guy, tante Ann, Peet, Bompa, Lara, tante Ludwine... Het is bijzonder fijn om altijd te kunnen vertrouwen op zo een geweldige familie!

Schattie, deze sectie is niet groot genoeg om jou te bedanken voor alles wat je voor me betekent, jou leren kennen is het beste dat me overkomen is tijdens mijn doctoraat. Niet alleen versta jij als geen ander de frustraties die gepaard gaan met een doctoraat, je bent daarenboven ook nog eens één van de weinige mensen die al mijn andere frustraties probeert te begrijpen. Danke! Je maakt me beter en laat me elke dag opnieuw inzien hoe geweldig Superteam is! Ich liebe dich.

Kim Hendrickx
31 Mei 2018
Diepenbeek

Samenvatting

Statistische modellen worden gebruikt om gegevens te analyseren en verbanden tussen geobserveerde variabelen te beschrijven. Voor sommige modellen is de onderstelling dat het verband tussen de variabelen monotoon is, een natuurlijke aanname. In dat geval moet deze monotoniciteitseis ingebouwd worden in de procedure die ontwikkeld wordt om de data te analyseren. In het eerste deel van deze thesis bespreken we statistische modellen waarin dergelijke monotoniciteitsvoorwaarden aan de orde zijn. De concrete problemen worden in detail beschreven in het introductie hoofdstuk van Deel 1.

In Hoofdstuk 2 zijn we geïnteresseerd in een lineair verband tussen een responsvariabele enerzijds en een groep covariaten anderzijds. De verklarende covariaatvariabelen worden volledig waargenomen maar de responsvariabele wordt niet geobserveerd. We weten enkel of de responsvariabele voor of na een geobserveerde censureringsvariabele ligt. Een corresponderende censureringsindicator geeft aan of de responsvariabele groter of kleiner is dan de censureringsvariabele. Dit censureringsmechanisme wordt het "current status" model genoemd. Het doel van Hoofdstuk 2 is het schatten van parameters die het lineair verband tussen de respons en de covariaten beschrijven op basis van observaties voor de censureringsvariabele, indicator en covariaten.

Een welgekende techniek om de parameters in een regressiemodel te schatten maakt gebruik van de “maximum likelihood” methode waarbij de schatters voor de regressieparameters gevonden worden door de aannemelijkheid van de data te maximaliseren. Het gecensureerde karakter van de observaties leidt tot een monotoniciteitseis voor een functie die voorkomt in de aannemelijkheidsfunctie van de data. Omtrent het gedrag van deze “maximum likelihood estimators" (MLEs) voor de regressieparameters in het current status model zijn nog een aantal vragen onopgelost in de statistische literatuur.

Op basis van de aannemelijkheidsfunctie construeren we in Hoofdstuk 2 een scorefunctie en bekomen we nieuwe schatters die gedefinieerd zijn via het nulpunt van deze scorefunctie. Onze schatters zijn daardoor nog steeds afhankelijkheid van het gedrag van de MLE. Door de studie van het asymptotisch gedrag van deze schatters, zijn we in staat om de vragen, nog steeds onopgelost voor de MLE, te beantwoorden voor de nieuwe scoreschatters die opgebouwd zijn vanuit het maximum likelihood mechanisme.

In Hoofdstuk 3 gaat onze interesse uit naar de verdelingsfunctie van de respons in afwezigheid van covariaten. De beschikbare informatie zijn de censureringsvariabelen en de censureringsindicatoren. De verdelingsfunctie bepaalt het gedrag van de responsvariabele en is per definitie monotoon stijgend. De bedoeling is om puntsgewijze betrouwbaarheidsintervallen te construeren voor de ongekende verdelingsfunctie. Opnieuw kan het maximum likelihood principe gebruikt worden om deze verdelingsfunctie te schatten (opnieuw MLE genoemd). Om betrouwbaarheidsintervallen op te stellen rond deze MLE voor de verdelingsfunctie, kan gebruik gemaakt worden van een bootstrapmethode die gebaseerd is op het genereren van gegevens door het hertrekken uit de geobserveerde waarnemingen. Uit de bestaande statistische literatuur is bekend dat de MLE in combinatie met de typische niet-parametrische bootstrapmethode geen goede betrouwbaarheidsintervallen oplevert. In Hoofdstuk 3 schatten we de verdelingsfunctie via een gladgemaakte versie van de MLE en tonen we aan dat de klassieke bootstrap wel gebruikt kan worden om betrouwbaarheidsintervallen op te stellen rond deze gladgemaakte MLE.

We stellen in dit hoofdstuk ook een tweede bootstrapalgoritme voor en vergelijken deze methode met de klassieke bootstrap. Het verschil tussen beide procedures is dat het klassieke hertrekken van censureringsvariabelen en indicatoren onafhankelijk van het onderliggende model kan uitgevoerd worden terwijl de tweede methode afhankelijk is van de gladgemaakte schatter voor de verdelingsfunctie en dat enkel indicatoren (en geen censureringsvariabelen) hertrokken worden.

Uit onze numerieke experimenten volgt er geen uitgesproken voorkeur voor één van beide bootstraptechnieken en het is opmerkelijk dat voor het construeren van betrouwbaarheidsintervallen voor de verdelingsfunctie, beide bootstrapmethoden het even goed lijken te doen. De extra informatie omtrent het onderliggende model bij de tweede bootstrapmethode leidt niet tot een verbetering van de niet-parametrische methode. De eenvoud van de klassieke procedure is een voordeel van de niet-parametrische bootstrapmethode. Een voordeel van de tweede bootstrap is dat enkel de indicatoren hertrokken worden, op die manier blijft de verdeling van de censureringsvariabelen ongewijzigd.

De techniek die gebruikt wordt om de oorspronkelijke MLE glad te maken, is afhankelijk van een parameter, die de bandbreedte genoemd wordt. In onze simulatiestudies bestuderen we het effect van deze bandbreedte op de betrouwbaarheidsintervallen. We illusteren hoe belangrijk een goede keuze voor deze parameter is en ontwikkelen een selectieprocedure voor de bandbreedte.

In het laatste hoofdstuk van Deel I bestuderen we het monotoon single index model. In dit regressiemodel worden zowel de covariaten als de respons volledig geobserveerd. De gegevens zijn niet gecensureerd. De respons hangt af van een lineaire combinatie van de covariaten via een ongekende linkfunctie. De enige voorwaarde op deze linkfunctie is dat ze een monotoon stijgend verloop heeft. Het doel is om de regressieparameters te schatten, dit zijn de parameters die het lineair verband in de covariaten bepalen. In de schattingsprocedure moet rekening gehouden worden met het stijgend karakter van de linkfunctie. We werken met kleinste kwadratenschatters (KKS). Deze minimaliseren de gekwadrateerde foutensom die ontstaat door de som te nemen van de gekwadrateerde verschillen tussen de geobserveerde respons en de voorspelde respons in het geschatte model.

De statistische eigenschappen van de KKS zijn vergelijkbaar met de eigenschappen van de MLE in het current status lineair regressie model van Hoofdstuk 2. Het asymptotisch gedrag van de KKS werd tot op heden nog niet helemaal doorgrond.

We construeren vanuit de gekwadrateerde foutensom een scorefunctie die leidt tot scoreschatters voor de regressieparameters in het monotoon single index model. Uit een vergelijking van deze scoreschatters met reeds eerder voorgestelde rangschatters in Han (1987) en in Cavanagh and Sherman (1998), volgt dat beide schatters weliswaar dezelfde asymptotische convergentiesnelheid hebben maar de scoreschatters hebben een kleinere variantie dan de rangschatters. Op basis van onze simulatie-experimenten kunnen we geen overtuigende conclusies trekken omtrent de asymptotische convergentiesnelheid van de KKS. We kunnen wel besluiten dat scoreschatters een kleinere variantie hebben in eindige steekproeven dan de KKS die beiden gebaseerd zijn op een kleinste kwadratenprocedure.

In het tweede deel van deze thesis, bestuderen we een model waarbij de regressiecoëfficiënten functies zijn van covariaten (varying coefficients model, VCM) en waarbij niet alle responsvariabelen geobserveerd worden door de aanwezigheid van rechtse censurering. De respons is dan voor een aantal observaties niet gekend, er wordt enkel een ondergrens waargenomen. We ontwikkelen twee datatransformatie technieken waardoor een nieuwe variabele met dezelfde verwachtingswaarde als de niet-geobserveerde respons geconstrueerd wordt. De getransformeerde variabele wordt gebruikt in een gepenaliseerde kleinste kwadratenprocedure waarbij spline schatters gebruikt worden om de onderliggende variërende coëfficiënten functies te schatten.

De theoretische eigenschappen van deze gepenaliseerde spline schatters maken de schatters vergelijkbaar met reeds eerder gedefinieerde schatters voor het VCM met rechts gecensureerde observaties. Aan de hand van simulatiestudies en data voorbeelden illustreren we de kwaliteit van de voorgestelde schattingsprocedure.

Summary

Statistical models are used to analyze data and to search for relationships between observed variables. In many models, shape constraints are imposed and therefore, the procedures that are developed to analyze the data need to include the shape restriction. In the first part of this thesis, we discuss statistical applications that have to take a monotonicity constraint into account. The specific problems are discussed in more detail in the Introduction of Part I.

In Chapter 2, we model a linear relationship between a response variable and a set of covariate variables. The covariates are fully observed but the response variable is subject to type 1 interval censoring. Instead of observing the response, a censoring variable is observed together with an indicator informing about whether or not the unobserved response lies before or after the censoring variable. This type of censored data is known as current status data. One could say that each observation indicates the *current status* of the response at the observed censored value. The objective of Chapter 2 is to estimate the regression parameters that describe the linear relationship in the covariates based on observations for the censoring variable, the indicator and the covariates.

A well-known technique to obtain the estimators in a regression model is the maximum likelihood approach, where the estimators are defined by the regression parameters that maximize the likelihood of the observed data. The censored nature of the data results in a monotonicity constraint for a function that appears in the likelihood function of the data. Since the eighties, researchers have investigated the behavior of this so-called maximum likelihood estimator (MLE) of the regression parameters in the current status linear regression model and a lot of open questions still exist.

We derive a score function from the likelihood function and develop estimators that are

defined by the root of this score function. Consequently, our score estimators still depend on the behavior of the MLE. Based on the asymptotic study of the score estimators, we are able to answer questions for the score estimators, that are still unsolved questions for the MLE. It is the first time that estimators for the regression parameters are developed that depend on the behavior of the maximum likelihood procedure and that converge at the parametric rate to the true regression parameters in the current status linear regression model.

In Chapter 3, we focus on estimating the distribution function of the response variable in absence of covariate information. The available data consists of censoring variables and censoring indicators. The distribution function, which completely defines the behavior of the response, is a monotone increasing function. The aim of Chapter 3 is to construct pointwise confidence intervals for this unknown distribution function. The principle of maximum likelihood can again be used to estimate the distribution function under a monotonicity constraint. We call this estimator the MLE of the distribution function. Confidence intervals can be centered around this MLE using a bootstrap procedure which consists of resampling data from the observed sample. It was proved in Abrevaya and Huang (2005) that a combination of the MLE with the classical nonparametric bootstrap proposed by Efron (1979) leads to incorrect confidence intervals for the distribution function under current status data. We propose to estimate the distribution function by a smoothed version of the MLE (the smoothed maximum likelihood estimator, SMLE) and show that the nonparametric bootstrap does result in valid confidence intervals around the SMLE.

We also propose a second model-based bootstrap procedure that depends on the SMLE. In this procedure, the censoring variables in the bootstrap sample are the same and only the censoring indicators are resampled. In the nonparameteric bootstrap algorithm, both censoring variables and indicators are resampled with replacement from the original observations, independent of the true underlying model. We compare the quality of the two bootstrap procedures for constructing confidence intervals for the distribution function under current status data.

In our numerical experiments, it is not clear which of two bootstrap procedures is better and the most striking finding is the similarity of the results between the smooth and non-parametric bootstrap. The additional information on the underlying model in the smooth bootstrap method does not result in an improvement of the nonparametric bootstrap for the construction of pointwise confidence intervals for the distribution function under current status data. An advantage of the purely nonparametric bootstrap is its conceptual simplicity. An advantage of the smooth bootstrap is that only indicators are resampled and that in this sense, one stays closest to the sampling distribution of the censoring variable.

The smoothing technique used to construct the SMLE from the MLE depends on a bandwidth parameter. In the simulation studies of Chapter 3, much attention is given to the effect of the bandwidth on the confidence intervals. We demonstrate the importance of a proper bandwidth choice and we develop a selection procedure for the bandwidth parameter that results in good confidence intervals for the distribution function.

In Chapter 4, we extend the findings of Chapter 2 for the current status linear regression model to the monotone single index model. Both covariates and response are fully observed in this regression model. The response depends on a linear combination of the covariates, i.e. the single index of the covariates, via an unknown link function. The only assumption that one makes for this link function is that it has a monotone increasing behavior. The goal of Chapter 4 is to estimate the regression parameters that describe the linear combination of the covariates and hence to determine the single index component in this model. The monotonicty constraint on the link function has to be taken into account during the estimation process.

We analyze the behavior of the least squares estimator (LSE). In this algorithm, we search for the regression parameters and the monotone increasing link function that minimize the sum of squared errors which arises by taking the sum of the squared difference between the observed responses and the predicted responses in the estimated model. The asymptotic properties of the LSE of the regression parameters in the monotone single index model are comparable to the properties of the MLE in the current status linear regression model and so far, the behavior of this LSE was not yet fully understood.

We derive a score function from the sum of squared errors and define score estimators by the root of this score function. From a comparison between the score estimators and the rank estimators proposed in Han (1987) and in Cavanagh and Sherman (1998), we conclude that both estimators have the same asymptotic convergence rate but the score estimators have smaller variances than the rank estimators. Based on our simulation experiments, we did not get conclusive insights into the converge rate of the LSE. However, even if the LSE has the same convergence rate as the score estimators, our findings do show a better finite sample behavior of the score estimators.

In Part II of the thesis, we look at a model where the regression coefficients are functions of the covariates (varying coefficient model, VCM) and where the observed responses are subject to random right censoring. For some observations, the response is unknown and only a lower bound is observed. We introduce two data-transformation

approaches that create a transformed variable which has the same expectation, conditionally on covariates, as the unobserved response variable. This transformed response variable is used in a penalized least squares procedure where we use splines to estimate the coefficient functions in the underlying VCM, referred to as P-spline estimates of the coefficient functions in the VCM. Our theoretical results and our simulations illustrate the quality of our proposed techniques for estimating a VCM subject to random right censoring. We also compare our estimates to the estimates proposed in Yang et al. (2014) and moreover discuss how the finite sample performance of the estimates in Yang et al. (2014) can be improved.

Publications

This dissertation is based on the following publications and reports:

Balabdaoui, F., Groeneboom, P. and **Hendrickx, K.** (2018). Score estimation in the monotone single index model. *Submitted for publication*.

Groeneboom, P. and **Hendrickx, K.** (2017a). Current status linear regression. *Accepted for publication in Ann. Stat.*

Groeneboom, P. and **Hendrickx, K.** (2017b). The nonparametric bootstrap for the current status model. *Electron. J. Stat.* 11(2):3446-3484.

Groeneboom, P. and **Hendrickx, K.** (2018a). Confidence intervals for the current status model. *Scand. J. Stat.,* 45(1):135-163.

Groeneboom, P. and **Hendrickx, K.** (2018b). Estimation in monotone single index models. *Stat. Neerl.* doi:10.1111/stan.12138.

Hendrickx, K., Janssen, P. and Verhasselt, A. (2017). Penalized spline estimation in varying coefficient models with censored data. *TEST.* doi:10.1007/s11749-017-0574-y.

Contents

List of Tables

List of Figures

List of Abbreviations

Here, we give a list of the most often used abbreviations in the thesis.

$\xrightarrow{p}$	:	Convergence in probability
$\xrightarrow{d}$	:	Convergence in distribution
$\|\cdot\|$	:	Eucledian norm
CI	:	Confidence Interval
ESE	:	Efficient Score Estimator
GCM	:	Greatest Convex Minorant
i.i.d.	:	independent and identically distributed
LCM	:	Least Concave Majorant
LR	:	Likelihood Ratio
LSE	:	Least Squares Estimator
MLE	:	Maximum Likelihood Estimator
MSE	:	Mean Squared Error
SBF	:	Smooth Backfitting
SMLE	:	Smoothed Maximum Likelihood Estimator
SIM	:	Single Index Model
SSE	:	Simple Score Estimator
VCM	:	Varying Coefficient Model

Part I

Shape constrained estimation in the current status and single index model

Chapter 1

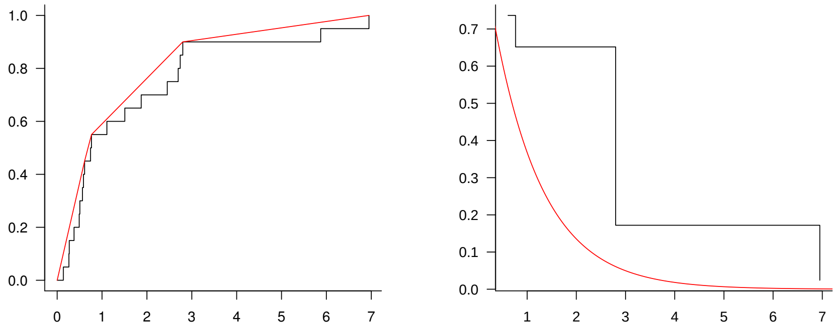
Introduction

Several statistical applications are based on imposing constraints that occur from the problem under study: monotonicity, convexity or concavity constraints arise naturally with consumption or production functions, growth curves and dose response models.Therefore, algorithms for shape constrained regression or density estimation are needed. Research on nonparametric estimation under shape constraints dates back to the 1950s and the papers such as the ones by Ayer et al. (1955) and van Eeden (1956) on estimation of functions under the constraint of monotonicity or unimodality. The classical example of estimating a monotone decreasing density f_0 on $[0, \infty)$ based on a sample $X_1, \ldots, X_n$ of i.i.d. observations from the unknown density was considered in Grenander (1956). The Maximum Likelihood Estimator (MLE) $\hat{f}_n$ that maximizes

$$f \mapsto \sum_{i=1}^{n} \log f(X_i),$$

over all decreasing densities $f : [0, \infty) \to [0, \infty)$, is the left derivative of the least concave majorant (LCM) of the empirical distribution function, where the LCM of a function $g : [0, \infty) \to \mathbb{R}$ is the lowest concave function that lies above g. The estimator is referred to as the *Grenander estimator*. For a sample of size $n = 20$ from the standard exponential distribution, the empirical distribution function together with its LCM are shown in Figure 1.1a. The resulting estimator $\hat{f}_n$ is shown in Figure 1.1b. It was proved in Prakasa Rao (1969) that the pointwise limit distribution of the Grenander estimator has a nonstandard limit distribution.

Theorem 1.0.1 (Prakasa Rao, 1969)**.** *Let $\hat{f}_n$ be the Grenander estimator of the density f_0 under the monotonicity restriction. Suppose that f_0 has a strictly negative derivative*



(a) **(b)**

Figure 1.1: Left: The empirical distribution function (black, step-function) and its LCM (red, solid). Right: The *Grenander estimator* (black, step function) and the standard exponential density (red, solid) for a sample of size $n = 20$ from the standard exponential distribution.

f_0' at the interior point x and $f_0(x) > 0$. Then

$$n^{1/3}\{\hat{f}_n(x) - f_0(x)\}/|4f_0(x)f_0'(x)| \xrightarrow{d} Z \qquad n \to \infty,$$

where $\xrightarrow{d}$ denotes convergence in distribution and $Z = \arg\max_t\{W(t) - t^2\}$, where W is the two-sided Brownian motion on $\mathbb{R}$ starting at zero.

The distribution of the random variable Z in Theorem 1.0.1 is known as the *Chernoff distribution*.

The behavior of estimators of increasing functions can often be derived similarly to the behavior of decreasing functions but involves greatest convex minorants (GCMs) instead of LCMs. As an example, we consider estimating the isotonic regression model based on observations $(x_i, y_i), 1 \leq i \leq n$, where the x_i are fixed and the y_i are realizations of the random variable Y_i with mean $\mu_0(x_i)$, for some increasing function $\mu_0 : \mathbb{R} \to \mathbb{R}$. Denote by $\bar{y}_k$ the average of the observations with mean $\mu_0(x_{(k)})$, where $x_{(1)} < \ldots < x_{(m)}$ are the ordered design points and m is the number of distinct design points. The Least Squares Estimator (LSE) $\hat{\mu}_n$ minimizes the weighted sum of squares

$$\mu \mapsto \sum_{k=1}^{m} w_k\{\bar{y}_k - \mu(x_{(k)})\}^2, \tag{1.0.1}$$

where $w_k, 1 \leq k \leq m$ are the number of observations with mean $\mu_0(x_{(k)})$, over all increasing functions $\mu : \mathbb{R} \mapsto \mathbb{R}$. The LSE $\hat{\mu}_n$ has the following characterization. Define the cumulative sum diagram consisting of the points $P_i = \left(\sum_{j=1}^{i} w_j, \sum_{j=1}^{i} w_j\bar{y}_j\right), 1 \leq i \leq m$

and the point $P_0 = (0, 0)$. Let $\hat{M}$ be the greatest convex function on $[0, n]$ lying completely below the set of points $\{P_i : 0 \leq i \leq m\}$. Then the LSE $\hat{\mu}_n(x_{(i)})$ is the left derivative of $\hat{M}$ evaluated at the point P_i. Since the sum of squares in (1.0.1) only depends on the values of the function μ evaluated at the observed design points, changing μ between two observed $x_{(i)}$ does not change the value of the objective function. We therefore consider functions μ that are piecewise constant between two successive design points.

In the first part of this thesis, we address two different models that involve a monotonicity constraint. We first address the current status model where the monotonicity constraint arises in the expression of the likelihood of the observed data. In the second model, we look at the monotone single index model. In this regression setting, a response depends on a linear combination of covariate variables via a link function that is assumed to be monotone. An introduction to both models is given in Section 1.1 and Section 1.2 respectively.

1.1 The current status model

Survival models are commonly used to characterize the distribution of a variable Y that is not observed directly. Depending on what information is obtained on Y, different censoring schemes arise. In the current status model, the variable Y of interest is only known to lie before or after some random censoring variable T. Each observed sample consists of a set of n inspection times T_i (independent of the other T_j and all $Y_j's$) and n censoring indicators $\Delta_i = 1_{\{Y_i \leq T_i\}}$. One could say that the ith observation indicates the *current status* of component i at time T_i. This type of censoring is also known as type I interval censoring and arises naturally in reliability and survival studies when the status of an observational unit is only checked at one measurement point, which happens especially when testing is destructive. For example, in carcinogenicity experiments, one is interested in the time from exposure to a potential carcinogen until the development of the tumor (Finkelstein and Wolfe, 1985 and Finkelstein, 1986). The presence of the tumor can, however, only be determined after animal sacrifice. Current status data is also obtained in epidemiological studies where the age of incidence of a certain disease is of interest (Keiding, 1991 and Keiding et al., 1996). The exact age at which the disease occurred is unknown, but a diagnostic test can be used to detect the presence of antibodies and hence inform about whether or not the disease already occurred in the past, at least if humoral immunity is preserved for life after infection.

For this type of data, the log likelihood function for estimating the distribution function F (ignoring a term that does not depend on F) is given by

$$\ell_n(F) = \sum_{i=1}^{n} \Delta_i \log F(T_i) + (1-\Delta_i)\log\{1-F(T_i)\}. \tag{1.1.1}$$

The MLE of the true distribution function F_0 of Y maximizes ℓ_n over all possible distribution functions F, i.e.

$$\hat{F}_n \stackrel{\text{def}}{=} \arg\max_F \ell_n(F). \tag{1.1.2}$$

Let $w_j, 1 \le j \le m$ be the weights given by the number of observations at point $T_{(j)}$, assuming that $T_{(1)} < \cdots < T_{(m)}$ are the m order statistics of the sample $T_1, \ldots, T_n$ where m is the number of different observations in the sample. Suppose that f_{1j} is the number of Δ_k equal to one at the jth order statistic of the sample. The value of $\hat{F}_n$ at $T_{(i)}$ can be characterized as the left continuous slope of the GCM (evaluated at the points P_i) of the cumulative sum diagram formed by the points $P_0 = (0,0)$ and

$$P_i = \left(\sum_{j=1}^{i} w_j, \sum_{j=1}^{i} f_{1j}\right), \; 1 \le i \le m.$$

In this thesis we address two different problems related to current status data.

1. The objective in Chapter 2 is to estimate a regression parameter in a linear regression model if on top of observing $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$ also covariate information for Y is available.

2. In Chapter 3 we investigate the validity of different bootstrap schemes for producing pointwise confidence intervals for the distribution function under current status data.

Both topics are discussed in more detail in Section 1.1.1 and Section 1.1.2.

1.1.1 Semiparametric regression under current status data

Instead of the completely nonparametric model described above, one can also consider the semiparametric current status regression model where one observes a sample $(\boldsymbol{X}_1, T_1, \Delta_1), \ldots, (\boldsymbol{X}_n, T_n, \Delta_n)$ from the random vector $(\boldsymbol{X}, T, \Delta)$ where $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is a $d-$dimensional covariate vector and (T, Δ) is defined as above. The unobserved response variable Y is modeled by

$$Y = \boldsymbol{\alpha}_0^T \boldsymbol{X} + \varepsilon, \tag{1.1.3}$$

with $\boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0d})^T$ being the $d-$dimensional regression parameter of interest and with ε being an unobserved random error term independent of T and $\boldsymbol{X}$ with unknown distribution function F_0. Although that the response variable Y in survival studies is mostly positive, we do not assume that the response Y is a nonnegative random variable in model (1.1.3) and our estimation techniques are also applicable to negative response variables (see e.g. Murphy et al. (1999) for a similar definition in the current status linear regression model).

In the semiparametric current status model, the distribution F_0 is no longer the parameter of interest but merely a nuisance parameter. Hence, model (1.1.3) is parametrized by the finite dimensional regression parameter $\boldsymbol{\alpha}_0 \in \mathbb{R}^d$ and the infinite dimensional nuisance parameter F_0. The relevant part of the log likelihood for estimating $(\boldsymbol{\alpha}_0, F_0)$ is given by

$$\sum_{i=1}^{n} \left[\Delta_i \log F_0(T_i - \boldsymbol{\alpha}_0^T \boldsymbol{X}_i) + (1 - \Delta_i) \log\{1 - F_0(T_i - \boldsymbol{\alpha}_0^T \boldsymbol{X}_i)\}\right].$$

The difficulty of the model is that the parametric part is "inside" the nonparametric part. One has to bypass the nuisance function F_0, which cannot be estimated at the parametric $\sqrt{n}$-rate, to get to the parametric part. The phenomenon is referred to as the bundled parameters problem in Ding and Nan (2011). This is very different for the Cox proportional hazards model for current status data defined by

$$\lambda(t|\boldsymbol{X}) = \lim_{dt \to 0} P(t < Y < t + dt | Y > t, \boldsymbol{X})/dt = \lambda_0(t) \exp(\boldsymbol{\alpha}_0^T \boldsymbol{X}),$$

where Y is assumed to be a nonnegative random variable, for some baseline hazard function λ_0. In this case, the log likelihood is of the form

$$\sum_{i=1}^{n} \left\{\Delta_i \log\left(1 - \exp\left\{-\Lambda_0(T_i) \exp(\boldsymbol{\alpha}_0^T \boldsymbol{X}_i)\right\}\right) - (1 - \Delta_i)\Lambda_0(T_i) \exp(\boldsymbol{\alpha}_0^T \boldsymbol{X}_i)\right\},$$

where Λ_0 is the baseline cumulative hazard function given by $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$. Now, the regression parameter $\boldsymbol{\alpha}_0$ does not appear in the argument of a function which is not $\sqrt{n}$-estimable and we can estimate Λ_0 and $\boldsymbol{\alpha}_0$ separately. In this case, it was shown in Huang (1996) that one can use the nonparametric MLE of Λ_0 and then use profile likelihood to estimate $\boldsymbol{\alpha}_0$ efficiently at $\sqrt{n}-$rate. However, for the ordinary current status regression model, it is still unknown whether a similar estimation method gives a $\sqrt{n}$-consistent estimate of $\boldsymbol{\alpha}_0$. The profile MLE of $\boldsymbol{\alpha}_0$ in model (1.1.3) was proved to be consistent in Cosslett (1983) and an $n^{1/3}-$rate (cube-root n) of convergence was derived in Murphy et al. (1999), but nothing is known about the asymptotic distribution of this profile MLE.

The efficient score function for $\boldsymbol{\alpha}_0$ is given by

$$\tilde{\ell}_{\boldsymbol{\alpha}_0,F_0}(\boldsymbol{x},t,\delta) = \left\{\mathbb{E}(\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X} = t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - \boldsymbol{x}\right\} f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) \cdot \left\{\frac{\delta}{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})} - \frac{1-\delta}{1-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})}\right\}, \qquad (1.1.4)$$

where $\mathbb{E}(\boldsymbol{Z}_1|\boldsymbol{Z}_2 = \boldsymbol{z}_2)$ denotes the conditional expectation of $\boldsymbol{Z}_1$ given $\boldsymbol{Z}_2 = \boldsymbol{z}_2$ for random vectors $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ and $f_0 = F_0'$ is the density of the underlying model (see Cosslett (1987), Huang and Wellner (1993) and Murphy et al. (1999) among others). $\sqrt{n}$-consistent estimators of $\boldsymbol{\alpha}_0$ with asymptotic normal limiting distribution where n times the variance equals the inverse of

$$\mathbb{E}\left\{\tilde{\ell}_{\boldsymbol{\alpha}_0,F_0}(\boldsymbol{X},T,\Delta)\tilde{\ell}_{\boldsymbol{\alpha}_0,F_0}(\boldsymbol{X},T,\Delta)^T\right\},$$

are called efficient estimators of $\boldsymbol{\alpha}_0$.
Approaches of efficient estimation of $\boldsymbol{\alpha}_0$ are considered in Murphy et al. (1999) (for the one-dimensional case) and in Li and Zhang (1998) and Shen (2000) among others. Proving the asymptotic normal distribution of an efficient estimator is often complicated due to the fact that the factors $F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})$ and $1-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})$ appear in the denominator of the efficient score function $\tilde{\ell}_{\boldsymbol{\alpha}_0,F_0}$. To avoid division by zero in some technical parts of the proofs, additional assumptions on the underlying model or truncation techniques in the estimation algorithm are needed. The authors of Murphy et al. (1999) and Shen (2000) assume that the data only provide information about a part of the distribution function F_0. This results in the condition that the support of the density of the random variable $T-\boldsymbol{\alpha}^T\boldsymbol{X}$ is strictly contained in an interval D and that F_0 stays away from zero and one on D for all $\boldsymbol{\alpha}$ in the parameter space. The drawback of the assumption is that we have no information about the whole distribution function F_0. It also goes against the usual conditions made for the current status model, where one commonly assumes that the observations provide information over the whole range of the distribution one wants to estimate. We presume that this assumption is made, among others, to avoid truncation devices that can prevent the problems arising if this condition is not made, such as unbounded score functions and numerical difficulties. Examples of truncation methods can be found in Cosslett (2007) and Klein and Spady (1993) to name a few, where the authors consider truncation sequences that converge to zero with increasing sample size.

All efficient estimates proposed in the semiparametric current status linear regression model are based on smoothing techniques in the estimation procedure. Although smoothing is unavoidable to obtain efficiency, it is not necessary to achieve the parametric $\sqrt{n}$-rate of convergence. Sherman (1993) showed that the maximum rank correlation estimator proposed by Han (1987) is a $\sqrt{n}$-consistent and asymptotically normal estimator. The estimator is motivated by a correlation argument and does not use an estimate of the underlying distribution function F_0. In Chapter 2, we construct a $\sqrt{n}$-consistent estimator of $\boldsymbol{\alpha}_0$ that is based on the nonparametric MLE of the underlying distribution function F_0. As far as we know, this is the first time that a non-smooth estimator of F_0 is used for constructing $\sqrt{n}-$consistent estimators.

1.1.2 Does the bootstrap work?

Since the introduction of the bootstrap approach by Efron (1979), resampling techniques have gained a lot of popularity for doing inference about a population based on a random sample. However, under current status data, several negative results have been published on the use of the bootstrap for generating the limit distribution of the MLE $\hat{F}_n$ of the distribution function F_0, defined in (1.1.2) (Abrevaya and Huang, 2005, Sen and Xu, 2015). From Groeneboom and Wellner (1992) we have the following result for the limiting distribution of the MLE $\hat{F}_n$:

Theorem 1.1.1 (Groeneboom and Wellner, 1992)**.** *Consider the current status model and let t be such that $0 < F_0(t), G(t) < 1$, and let F_0 and G be differentiable at t with strictly positive derivatives $f_0(t)$ and $g(t)$, respectively. Here, G denotes the distribution function of the censoring variable T. Furthermore, let $\hat{F}_n$ be the MLE of the distribution function F_0 of Y. Then, we have, as $n \to \infty$*

$$n^{1/3}\left\{\hat{F}_n(t) - F_0(t)\right\} \xrightarrow{d} [4F_0(t)\{1 - F_0(t)\}f_0(t)/g(t)]^{1/3}Z,$$

where $Z = \arg\max_t\{W(t) - t^2\}$ and $W(t)$ is a standard two-sided Brownian motion process, originating from zero.

For Efron's nonparametric bootstrap procedure, which consists of resampling n pairs (T_i, Δ_i) with replacement from the original sample $(T_1, \Delta_1), \dots, (T_n, \Delta_n)$, it follows from Abrevaya and Huang (2005) that (conditional on the data)

$$\begin{aligned} &n^{1/3}\{4F_0(t)(1 - F_0(t))f_0(t)/g(t)\}^{-1/3}\{\hat{F}_n^*(t) - \hat{F}_n(t)\} \\ &\quad \xrightarrow{d} \quad \arg\max_t(W(t) + \hat{W}(t) - t^2) - \arg\max_t(W(t) - t^2), \end{aligned} \tag{1.1.5}$$

where $\hat{F}_n^*$ is the bootstrap MLE, obtained by maximizing the log likelihood defined by (1.1.1) but with (T_i, Δ_i) replaced by the bootstrap observations, and where W and $\hat{W}$ are two independent two-sided Brownian motions originating at zero. A similar result is obtained in Kosorok (2008) for the Grenander estimator and in Abrevaya and Huang

(2005) for the maximum score estimator of Manski (1975), which is another example of a cube-root n statistic (Kim and Pollard, 1990). These results imply that Efron's bootstrap cannot generate the correct limiting distribution of the corresponding estimators and hence leads to inconsistent bootstrap-based confidence intervals.

Constructing asymptotic confidence intervals for the distribution function in the current status model based on Chernoff's distribution, defined in Theorem 1.0.1, and the normalizing constant $4F_0(t)\{1-F_0(t)\}f_0(t)/g(t)$ is complicated by the need to compute the critical values of Z and to estimate the density f_0 consistently. Accurate density estimation turns out to be a rather difficult task when relying on current status data and several alternative methods have been proposed.

Banerjee and Wellner (2005) came up with a likelihood-ratio (LR) based method. Starting from the likelihood ratio statistic

$$LR(\theta_0) = 2\left(\log \ell_n(\hat{F}_n) - \log \ell_n(\hat{F}_n^{\theta_0})\right),$$

for testing the null hypothesis $F_0(t) = \theta_0$, which has asymptotic distribution $\mathbb{D}$ characterized in Banerjee and Wellner (2001), the authors estimate the interval by

$$\{\theta \in (0,1) : LR(\theta) \le d_{1-\alpha}\},$$

where $d_{1-\alpha}$ is the $(1-\alpha)$th percentile of $\mathbb{D}$. Here $\hat{F}_n$ denotes the unconstrained MLE and $\hat{F}_n^{\theta_0}$ denotes the MLE of F_0, maximizing the log likelihood under the constraint that $F_0(t) = \theta_0$. The LR-based method avoids estimation of f_0 and g since, under the null hypothesis, the limiting distribution $\mathbb{D}$ does not depend on the underlying parameters.

Instead of Efron's bootstrap, Sen and Xu (2015) introduced a different model-based resampling scheme and proved the consistency of their bootstrap for constructing pointwise confidence intervals around the MLE $\hat{F}_n$, under certain smoothness conditions. The bootstrap algorithm will be referred to as the *smooth bootstrap* in this thesis and works as follows:

1. Obtain an initial estimate $\tilde{F}$ of F_0.
2. Generate censoring indicators $\Delta_i^*, 1 \le i \le n$ from a Bernoulli($\tilde{F}(T_i)$) distribution where the $T_i, 1 \le i \le n$ are kept the same as in the original sample $(T_1, \Delta_1), \dots (T_n, \Delta_n)$.
3. Consider the bootstrap sample $(T_1, \Delta_1^*), \dots (T_n, \Delta_n^*)$.

Consequently, pointwise confidence intervals can be formed by taking

$$\left[\hat{F}_n(t) - V^*_{1-\alpha/2}(t), \hat{F}_n(t) - V^*_{\alpha/2}(t)\right],$$

where V^*_α is the αth quantile of B values of

$$\hat{F}^*_n(t) - \tilde{F}(t),$$

where B is the number of bootstrap samples taken, and $\hat{F}^*_n(t)$ is the MLE in the bootstrap sample $(T_1, \Delta^*_1), \ldots (T_n, \Delta^*_n)$. The consistency of this method requires the initial estimate $\tilde{F}$ to satisfy the following smoothness condition for a point t_0 in the interior of the support of F_0

$$\lim_{n\to\infty} |\tilde{F}(t_0 + n^{-1/3}t) - \tilde{F}(t_0) - f_0(t_0)n^{-1/3}t| = 0 \qquad a.s.,$$

which is not satisfied by the MLE $\hat{F}_n$. Several other bootstrap results for the Grenander estimator can be found in the literature where similar issues with the lack of smoothness of the function from which the bootstrap samples are drawn, are illustrated. The inconsistency of generating Chernoff's limiting distribution using bootstrap samples from the empirical distribution function or its LCM is proved in Sen et al. (2010). The authors suggest other methods for obtaining the cube-root n-consistency, either by smoothing or by using m out of n subsampling. Kosorok (2008) shows that generating bootstrap samples from a smoothed Grenander estimator can give consistent L_1-confidence bands for the Grenander estimator.

The Smoothed Maximum Likelihood Estimator (SMLE, Groeneboom et al., 2010), obtained by first estimating the MLE and then smoothing this using a kernel, is an estimate of the distribution function that does result in consistent confidence intervals via the smooth bootstrap algorithm. The SMLE is defined as

$$\tilde{F}_{nh}(t) \stackrel{\text{def}}{=} \int \mathbb{K}\left((t-x)/h\right)\, d\hat{F}_n(x),$$

where $\mathbb{K}$ is the integrated kernel

$$\mathbb{K}(u) \stackrel{\text{def}}{=} \int_{-\infty}^{u} K(x)\, dx, \tag{1.1.6}$$

and where h is a chosen bandwidth. Here $d\hat{F}_n$ represents the jumps of the discrete distribution function $\hat{F}_n$ and K is a symmetric and twice continuously differentiable kernel function with compact support [-1,1]. For a constant $c > 0$ and $h = cn^{-1/5}$, the SMLE has been proved to converge at rate $n^{2/5}$ to a normal limit distribution.

Theorem 1.1.2 (Groeneboom et al., 2010)**.** *Let the distribution of F_0 have support $[0, M]$ and let F_0 have a density f_0 staying away from zero on $(0, M)$. Furthermore, let G have a density g with a support that contains $[0, M]$ and let g stay away from zero on $[0, M]$, with a bounded derivative g'. Finally, let t be an interior point of $[0, M]$ such that f_0 has a continuous derivative f_0' at t. Then, if $h = cn^{-1/5}$,*

$$n^{2/5}\left\{\tilde{F}_{nh}(t) - F_0(t)\right\} \xrightarrow{d} N(\beta, \sigma^2),$$

where

$$\beta = \frac{c^2 f_0'(t)}{2}\int u^2 K(u)\,du \quad \text{and} \quad \sigma^2 = \frac{F_0(t)\{1 - F_0(t)\}}{cg(t)}\int K(u)^2\,du.$$

In light of Theorem 1.1.2, a drawback of the approach proposed by Sen and Xu (2015) is the fact that smoothness conditions of F_0 are used which allow faster than cube-root n estimation of F_0. This raises the question whether one should really use pointwise confidence intervals based on the MLE instead of on a faster converging estimate such as the SMLE. In Chapter 3, we address this question and show that both Efron's nonparametric bootstrap and the model-based bootstrap of Sen and Xu (2015) can be used to generate the asymptotic normal distribution of the SMLE. Nevertheless, for current status and related models, some research has been reported recommending the use of the smooth bootstrap procedure. A smooth bootstrap calibration was used in Durot and Reboul (2010) for a goodness-of-fit-test for monotone functions and in Groeneboom (2012) for a likelihood ratio type two-sample test for current status data. Durot et al. (2013) used a similar approach to determine the critical value for testing equality of functions under monotonicity constraints. The main motivation for recommending the smooth bootstrap were the negative results by Abrevaya and Huang (2005) and Kosorok (2008) proving the inconsistency of Efron's nonparametric bootstrap for generating the limiting distribution of the MLE. Although Durot and Reboul (2010) conjecture that the nonparametric bootstrap fails in their setting, the results presented in this thesis, however, suggest that this conjecture might be incorrect and that applications of the nonparametric bootstrap involving the Grenander estimator are worth studying in further research.

Besides considering the nonparametric or smooth bootstrap, one could moreover consider resampling the Δ_i from the MLE itself. Simulation studies in Durot et al. (2013) for testing equality of functions under monotonicity constraints even suggest that the smooth bootstrap does not necessarily perform better than bootstrapping from the non-smooth functional itself for certain smooth functionals different from the non-smooth functional $F \mapsto F(t)$. (See van der Vaart (1991) for the non-differentiability of the evaluation mapping $F \mapsto F(t)$ in the current status model. A similar result holds for the mapping

$f \mapsto f(t)$ in the estimation of a decreasing density). So far, the theoretical properties of the latter bootstrap procedure remain an open problem. As a consequence of our positive result on Efron's bootstrap (see Chapter 3), we conjecture that bootstrapping from the MLE might as well work for constructing pointwise confidence intervals in the current status model, if one uses the right (smooth) functional of the model as a basis for the intervals.

1.2 The single index model

Single index models (SIMs) are flexible semiparametric regression models used to investigate the relationship between a response variable Y and a covariate vector $\boldsymbol{X} = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$. The SIM is given by

$$Y = \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X}) + \varepsilon, \qquad (1.2.1)$$

where $\varepsilon \sim F_0$ is a random error term satisfying $\mathbb{E}(\varepsilon|\boldsymbol{X}) = 0$ and ψ_0 is an unknown link function. These models are more flexible than standard linear regression models and have, on the other hand, more structure than completely nonparametric models. By lowering the dimensionality of the classical linear regression problem, determined by the number of covariates, to a univariate $\boldsymbol{\alpha}_0^T \boldsymbol{X}$ index, SIMs do not suffer from the "curse of dimensionality". They also provide an advantage over the generalized linear regression models by overcoming the risk of misspecifying the link function ψ_0.

Identifiability of the single index regression parameter $\boldsymbol{\alpha}_0$ (up to a scalar constant) has been discussed in Ichimura (1993) in terms of the distribution of the regressors $\boldsymbol{X}$. Without any further restrictions, the parameter vector $(\boldsymbol{\alpha}_0, \psi_0)$ can, however, not be estimated in the SIM. This can be seen as follows. Take $a, b \in \mathbb{R}$ and let ψ^* be the function defined by the relationship $\psi^*(a + bt) = \psi_0(t)$ for all t in the support of $\boldsymbol{\alpha}_0^T \boldsymbol{X}$, then

$$\mathbb{E}(Y|\boldsymbol{X}) = \psi^*(a + b\boldsymbol{\alpha}_0^T \boldsymbol{X}).$$

Even if the distribution of $(\boldsymbol{X}, Y)$ is known, the above model cannot be distinguished from model (1.2.1) unless restrictions on the location a and scale b are imposed. Location normalization can be imposed by requiring that all components of $\boldsymbol{X}$ have a nondegenerate distribution. A reparametrization of the parameter space to the set

$$\{\boldsymbol{\alpha} \in \mathbb{R}^d : \alpha_1 = 1\} \quad \text{or} \quad \{\boldsymbol{\alpha} \in \mathbb{R}^d : \|\boldsymbol{\alpha}\| = 1, \alpha_1 \geq 0\},$$

where $\|\cdot\|$ denotes the Euclidean norm and α_1 is the first component of $\boldsymbol{\alpha}$, ensures scale identification of the model. The first parametrization is used in Sherman (1993), examples of the second parametrization are found in Härdle et al. (1993) and Hristache et al. (2001) among others.

The log likelihood of the model is given by

$$\sum_{i=1}^{n} \log \left\{ f_{\epsilon|\boldsymbol{X}} \left(Y_i - \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X}_i) \right) g(\boldsymbol{X}_i) \right\},$$

where $f_{\epsilon|\boldsymbol{X}}(\cdot|\boldsymbol{x})$ is the conditional density of ε given $\boldsymbol{X} = \boldsymbol{x}$ and g is the density of $\boldsymbol{X}$. The efficient score function for $\boldsymbol{\alpha}_0$ is given by

$$\tilde{\ell}_{\boldsymbol{\alpha}_0,\psi_0}(\boldsymbol{x}, y) = \frac{y - \psi_0\left(\boldsymbol{\alpha}_0^T \boldsymbol{x}\right)}{\sigma^2(\boldsymbol{x})} \psi_0'\left(\boldsymbol{\alpha}_0^T \boldsymbol{x}\right) \left\{ \boldsymbol{x} - \frac{\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})\boldsymbol{X}|\boldsymbol{\alpha}_0^T \boldsymbol{X} = \boldsymbol{\alpha}_0^T \boldsymbol{x}\right\}}{\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})|\boldsymbol{\alpha}_0^T \boldsymbol{X} = \boldsymbol{\alpha}_0^T \boldsymbol{x}\right\}} \right\},$$

where $\sigma^2(\cdot) = \mathbb{E}(\varepsilon^2|\boldsymbol{X} = \cdot)$ (see e.g. Delecroix et al., 2003 and Kuchibhotla and Patra, 2017). In the special case that $\varepsilon|\boldsymbol{X} \sim N(0, \sigma^2)$, the log likelihood of the model is, up to a term not depending on $(\boldsymbol{\alpha}_0, \psi_0)$, given by $S_n(\boldsymbol{\alpha}_0, \psi_0)/(2\sigma^2)$, where S_n is the sum of squared errors

$$S_n(\boldsymbol{\alpha}, \psi) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \left\{ Y_i - \psi(\boldsymbol{\alpha}^T \boldsymbol{X}_i) \right\}^2.$$

The LSE $(\hat{\boldsymbol{\alpha}}_n, \hat{\psi}_n)$ in the SIM, defined by

$$(\hat{\boldsymbol{\alpha}}_n, \hat{\psi}_n) \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{\alpha},\psi} S_n(\boldsymbol{\alpha}, \psi), \tag{1.2.2}$$

therefore coincides with the MLE in the SIM with homoscedastic normal error terms.

Besides the LSE, several other estimators have been proposed in the literature that can be classified into different groups based on the estimation algorithm. Most estimators require a nonparametric estimator of ψ_0. Often smoothing procedures, such as kernel smoothers or spline functions are used to avoid discontinuous criterion functions. An example of this type is the (weighted) semiparametric least squares estimator (SLSE), which corresponds to minimizing the sum of squares $S_n(\boldsymbol{\alpha}, \tilde{\psi}_{\boldsymbol{\alpha},h})$ over $\boldsymbol{\alpha}$, where $\tilde{\psi}_{\boldsymbol{\alpha},h}$ is a kernel estimator of ψ_0 that depends on $\boldsymbol{\alpha}$ and a bandwidth h. Härdle et al. (1993) extended the SLSE by minimizing this sum of squares over $(\boldsymbol{\alpha}, h)$ simultaneously to obtain optimal smoothing. Instead of estimating ψ_0 by a kernel smoother, spline smoothing is considered in Yu and Ruppert (2002) and Antoniadis et al. (2004) among others. The average derivative estimator proposed by Hristache et al. (2001), results in direct estimation of the regression parameter $\boldsymbol{\alpha}_0$ and therefore avoids solving a hard optimization problem, which is often the case with M-estimators. The idea of the average derivative method is to estimate the gradient $\boldsymbol{\alpha}_0^T\psi_0'(\boldsymbol{\alpha}_0^T\boldsymbol{x})$ of the link function using local linear smoothing techniques. A similar approach is considered for the minimum average variance estimator proposed in Xia and Härdle (2006). Except for the Bayesian estimation method proposed by Antoniadis et al. (2004), all other methods discussed in this paragraph are proved to converge at $\sqrt{n}$-rate to the true regression parameter in the single index model.

Model (1.2.1), when ψ_0 is an unknown monotone function is also known as the monotone SIM and many econometric models, censored regression models as well as various duration models fit into this framework. Below we describe how the binary choice model and the current status linear regression model are special cases of the monotone SIM satisfying $\mathbb{E}(Y|\boldsymbol{X}) = \psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})$.

The binary choice model: A widely used econometric model is the binary choice model which is used to describe a choice probability based on one or more covariates. The model is given by

$$Y = \begin{cases} 1 & \text{if } \boldsymbol{\alpha}_0^T\boldsymbol{X} \geq \tilde{\varepsilon} \\ 0 & \text{else,} \end{cases}$$

where $\boldsymbol{\alpha}_0^T\boldsymbol{X}$ represents the utility score and $\tilde{\varepsilon}$ is the disturbance term, which is assumed to be independent of $\boldsymbol{X}$. The model can be used to predict the probability that a person decides to consume a certain good based on the characteristics of the person. The model is a special case of the SIM (1.2.1) with ψ_0 equal to the (unknown and monotone) distribution function F_0 of $\tilde{\varepsilon}$, since

$$\mathbb{E}\{Y|\boldsymbol{X}\} = P(Y=1|\boldsymbol{X}) = P(\tilde{\varepsilon} \leq \boldsymbol{\alpha}_0^T\boldsymbol{X}) = F_0(\boldsymbol{\alpha}_0^T\boldsymbol{X}).$$

The current status linear regression model: For the current status linear regression model described in Section 1.1.1, we have

$$\mathbb{E}\{\Delta|\boldsymbol{X},T\} = P(\Delta = 1|\boldsymbol{X},T) = F_0(T - \boldsymbol{\alpha}_0^T\boldsymbol{X}).$$

The model is therefore a special case of the monotone SIM with response $\tilde{Y} = \Delta$, covariate vector $\tilde{\boldsymbol{X}} = (T, \boldsymbol{X})^T$ and $\tilde{\boldsymbol{\alpha}}_0 = (1, \boldsymbol{\alpha}_0^T)$. Since the first component of the covariate vector corresponds to the censoring variable T with coefficient equal to one, the current status regression model is identified without further restrictions on the parameter space.

The asymptotic properties of the LSE, defined in (1.2.2) for the monotone SIM (Tanaka, 2008) are comparable to the properties of the MLE for the binary choice model (Cosslett,

1983) and the current status linear regression model (Murphy et al., 1999). All estimators converge at rate cube-root n, but their limiting distribution is still an open problem. Examples of M-estimators (that are not based on an estimate of ψ_0) are Manski's maximum rank estimator (Manski, 1975) for the binary choice model and the maximum rank correlation estimator proposed by Han (1987) and the rank estimators proposed by Cavanagh and Sherman (1998) for a more general generalized regression model under monotonicity constraints. In contrast to the other estimators, Manski's estimator converges somewhat disappointingly at the cube-root n-rate to a nonstandard limiting distribution instead of at the usual parametric $\sqrt{n}$-rate to a normal limiting distribution (Kim and Pollard, 1990). In Chapter 4, we extend the regression estimators for the current status linear regression model proposed in Chapter 2 to estimators for the more general SIM. The estimators are derived from a score approach corresponding to the least squares minimization problem. This is again the first time that $\sqrt{n}$-consistent estimators of the finite dimensional regression parameter in the SIM are constructed based on the monotone LSE of the link function ψ_0.

1.3 Outline of the thesis

This thesis is organized as follows. In Chapter 2, we develop $\sqrt{n}$-consistent and asymptotically normal estimates of the finite dimensional regression parameter in the current status linear regression model defined in (1.1.3). These estimates do not require any smoothing device and are based on MLEs of the infinite dimensional distribution function parameter. We next construct estimates, again only based on these MLEs, which are arbitrarily close to efficient estimates if the generalized Fisher information is finite. This type of efficiency is also derived under minimal conditions for estimates based on smooth non-monotone plug-in kernel estimates of the distribution function which are independent of the MLE of the distribution function. Instead of a maximization approach, where one obtains an estimate as the maximizer of a criterion function, we take a slightly different angle and define all our estimates by the root of a score function. Algorithms for computing the estimates and for selecting the bandwidth of the smooth estimates with a bootstrap method are provided. The research of Chapter 2 is published in Groeneboom and Hendrickx (2017a).

In Chapter 3, we introduce a new way of constructing pointwise confidence intervals for the distribution function in the current status model. The confidence intervals are based on the SMLE, using local smooth functional theory and normal limit distributions. Two bootstrap methods for constructing the latter intervals are discussed. The methods proposed by Banerjee and Wellner (2005) and Sen and Xu (2015) to construct confidence intervals, using the nonstandard limit distribution of the (restricted) MLE, are compared to our approach via simulations. This chapter covers the material presented in Groeneboom and Hendrickx (2017b) and in Groeneboom and Hendrickx (2018a). The method proposed in the first paper is implemented in the R package `curstatCI`.

Estimation of the regression parameter in the single index model with monotone link function is considered in Chapter 4. Using the ideas of Chapter 2, we develop $\sqrt{n}$-consistent and asymptotically normal estimates of the finite dimensional regression parameter which are based on the LSE of the monotone link function. We also review different "non-smooth" estimates that avoid the standard approach of using smoothing techniques. We illustrate our score approach via simulations and compare our method with other methods such as Han's maximum rank correlation estimate. Chapter 4 summarizes the methods discussed in Balabdaoui et al. (2018) and Groeneboom and Hendrickx (2018b).

Chapter 5 deals with P-spline smoothing techniques in a varying coefficient regression model when the response is subject to random right censoring. We introduce two data-transformation approaches to construct a synthetic response vector that is used in a penalized least-squares optimization problem. We prove the consistency and asymptotic normality of the P-spline estimates for a diverging number of knots and show by simulation studies and real data examples that the combination of a data-transformation for censored observations with P-spline smoothing leads to good estimates of the varying coefficient functions. Chapter 5 presents the results given in Hendrickx et al. (2017).

An overview of future research directions is given in Chapter 6. All the proofs and technical details needed to prove the results presented in Chapters 2-5 are given in the Appendix.

Chapter 2

Current status linear regression

Abstract

We construct $\sqrt{n}$-consistent and asymptotically normal estimates of the finite dimensional regression parameter in the current status linear regression model. The first estimate is obtained from the root of a score equation which is derived from the likelihood of the observed data and depends on the maximum likelihood estimator of the infinite dimensional parameter. This score estimate does not require any smoothing device and is the first $\sqrt{n}$-consistent estimates that is derived from an estimation algorithm that depends on the piecewise constant MLE of the distribution function.

We also show that these simple score estimates can be improved when a smoothing device is implemented to estimate the density of the error term in the current status model. Hence, we construct estimates, again only based on the MLEs of the distribution function, which are arbitrarily close to efficient estimates, if the generalized Fisher information is finite. This type of efficiency is also derived for a third estimate under minimal conditions for smooth non-monotone kernel estimates of the distribution function instead of the MLEs of the distribution function. Algorithms for computing the three estimates and for selecting the bandwidth of the second and third smooth efficient estimates with a bootstrap method are provided. The connection with results in the econometric literature is also made.

Our simulation results show that all three estimates perform well in finite samples. The second estimate, depending on the MLE of the distribution function, has a slightly better finite sample performance in large samples than the asymptotic equivalent third estimate, depending on a smooth kernel estimate of the distribution function. For the MLE-based estimates, neither the first nor the second estimate comes out as uniformly best in our simulations with small sample sizes, even though that the theoretical properties of the second estimate are better than the properties of the first estimate.

2.1 Model description

Let $(\boldsymbol{X}_i, T_i, \Delta_i), 1 \leq i \leq n,$ be independent and identically distributed observations from $(\boldsymbol{X}, T, \Delta) = ((X_1, \ldots, X_d)^T, T, 1_{\{Y \leq T\}})$. We assume that Y is modeled as

$$Y = \boldsymbol{\alpha}_0^T \boldsymbol{X} + \varepsilon, \tag{2.1.1}$$

where $\boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0d})^T$ is a d-dimensional regression parameter in the parameter space $\Theta \subset \mathbb{R}^d$ and ε is an unobserved random error, independent of $(\boldsymbol{X}, T)$ with unknown distribution function F_0 and $\mathbb{E}(\varepsilon) = \mu_0$. We assume that each component of $\boldsymbol{X}$ is nondegenerate and that the distribution of $(\boldsymbol{X}, T)$ does not depend on $(\boldsymbol{\alpha}_0, F_0)$ which implies that the relevant part of the log likelihood for estimating $(\boldsymbol{\alpha}_0, F_0)$ is given by

$$\begin{aligned}\ell_n(\boldsymbol{\alpha}, F) &= \sum_{i=1}^n \left[\Delta_i \log F(T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i) + (1 - \Delta_i) \log\{1 - F(T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i)\}\right] \\ &= \int \left[\delta \log F(t - \boldsymbol{\alpha}^T \boldsymbol{x}) + (1 - \delta) \log\{1 - F(t - \boldsymbol{\alpha}^T \boldsymbol{x})\}\right] \, d\mathbb{P}_n(\boldsymbol{x}, t, \delta),\end{aligned} \tag{2.1.2}$$

where $\mathbb{P}_n$ is the empirical distribution of the $(\boldsymbol{X}_i, T_i, \Delta_i), 1 \leq i \leq n$. We will denote the probability measure of $(\boldsymbol{X}, T, \Delta)$ by P_0. Since the efficient score function, defined in (1.1.4), contains the factors $F_0(t - \boldsymbol{\alpha}_0^T \boldsymbol{x})$ and $1 - F_0(t - \boldsymbol{\alpha}_0^T \boldsymbol{x})$ in the denominator, we introduce a truncated log likelihood to avoid problems arising from division by zero in some technical steps of the proofs. The truncated log likelihood $\ell_{\epsilon,n}(\boldsymbol{\alpha}, F)$ is defined by

$$\begin{aligned}\ell_{\epsilon,n}(\boldsymbol{\alpha}, F) \stackrel{\text{def}}{=} \int_{F(t - \boldsymbol{\alpha}^t \boldsymbol{x}) \in [\epsilon, 1-\epsilon]} &\left[\delta \log F(t - \boldsymbol{\alpha}^T \boldsymbol{x})\right. \\ &\left. + (1 - \delta) \log\left\{1 - F(t - \boldsymbol{\alpha}^T \boldsymbol{x})\right\}\right] d\mathbb{P}_n(\boldsymbol{x}, t, \delta),\end{aligned} \tag{2.1.3}$$

where $\epsilon \in (0, 1/2)$ is a truncation parameter. Analogously, let

$$\psi_{\epsilon,n}(\boldsymbol{\alpha}, F) \stackrel{\text{def}}{=} \int_{F(t - \boldsymbol{\alpha}^T \boldsymbol{x}) \in [\epsilon, 1-\epsilon]} \phi(\boldsymbol{x}, t, \delta)\{\delta - F(t - \boldsymbol{\alpha}^T \boldsymbol{x})\} \, d\mathbb{P}_n(\boldsymbol{x}, t, \delta),$$

define the truncated score function for some weight function ϕ. In this chapter, we consider estimates of $\boldsymbol{\alpha}_0$, derived by the idea of solving a score equation

$$\psi_{\epsilon,n}(\boldsymbol{\alpha}, \hat{F}_{\boldsymbol{\alpha}}) = 0,$$

where $\hat{F}_{\boldsymbol{\alpha}}$ is an estimate of F for fixed $\boldsymbol{\alpha}$. A motivation for this score approach is outlined in Section 2.2.

Before we continue, we introduce the simulation set-up that will be used in the remainder of this chapter to illustrate our estimation techniques. We consider sampling from the one-dimensional model $Y = \alpha_0 X + \varepsilon$ where the true regression parameter $\alpha_0 = 0.5$ and where X and T are independent uniform random variables on $[0, 2]$. The error term ε has density f_0 given by

$$f_0(e) = 384(e - 3/8)(5/8 - e),$$

if $e \in [3/8, 5/8]$ and zero else. Since the expectation of the random error variable equals $\mu_0 = 0.5$, our simulation model contains an intercept, i.e. $\mathbb{E}(Y|X) = 0.5 + 0.5X$.

2.2 Motivation for the score approach

Instead of defining an estimator as the maximizer of a certain criterion function, one could alternatively define the estimator as the root of a score function. If we consider the profile maximum likelihood method, we first fix $\boldsymbol{\alpha}$ and estimate F_0 by the maximizer $\hat{F}_{n,\boldsymbol{\alpha}}$ of the likelihood given in (2.1.2). Although $\hat{F}_{n,\boldsymbol{\alpha}}$ is not differentiable, we can still use this estimate of F_0 in the score approach. Smoothing techniques can be used in the score approach but are not necessary as we will see in the definition of our simple score estimator in Section 2.4.1. As far as we known, no $\sqrt{n}$-consistent argmax estimator of $\boldsymbol{\alpha}_0$ has been defined based on the MLE $\hat{F}_{n,\boldsymbol{\alpha}}$ of the distribution function.

The "canonical" approach to proof that argmax estimates of $\boldsymbol{\alpha}_0$ are $\sqrt{n}$-consistent has been provided in Sherman (1993), where the author gives sufficient conditions to prove that

$$\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\| = O_p(n^{-1/2}), \tag{2.2.1}$$

where $\hat{\boldsymbol{\alpha}}_n$ is defined by

$$\hat{\boldsymbol{\alpha}}_n \stackrel{\text{def}}{=} \text{argmax}\Gamma_n(\boldsymbol{\alpha}),$$

for some criterion function Γ_n. If Γ is the population equivalent of Γ_n, Theorem 1 in Sherman (1993) says that (2.2.1) is satisfied if,

(a) there exists a neighborhood N of $\boldsymbol{\alpha}_0$ and a constant $k > 0$ such that

$$\Gamma(\boldsymbol{\alpha}) - \Gamma(\boldsymbol{\alpha}_0) \leq -k\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|^2,$$

for $\boldsymbol{\alpha} \in N$, and

(b) uniformly over $o_p(1)$ neighborhoods of $\boldsymbol{\alpha}_0$,

$$\begin{aligned}&\Gamma_n(\boldsymbol{\alpha}) - \Gamma_n(\boldsymbol{\alpha}_0)\\ &= \Gamma(\boldsymbol{\alpha}) - \Gamma(\boldsymbol{\alpha}_0) + O_p\left(\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|/\sqrt{n}\right) + o_p\left(\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|^2\right) + O_p\left(n^{-1}\right).\end{aligned}$$

Moreover, assuming without loss of generality that $\boldsymbol{\alpha}_0 = \mathbf{0}, \Gamma(\boldsymbol{\alpha}_0) = \mathbf{0}$, and that

$$\Gamma_n(\boldsymbol{\alpha}) = -\tfrac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{V}\boldsymbol{\alpha} + n^{-1/2}\boldsymbol{\alpha}^T\boldsymbol{W}_n + o_p\left(n^{-1}\right),$$

where $\boldsymbol{V}$ is a positive definite matrix and $\boldsymbol{W}_n$ converges in distribution to a normal distribution, $\sqrt{n}\boldsymbol{\alpha}_n$ also converges to a normal distribution.

If we try to apply this to the MLE of $\boldsymbol{\alpha}_0$, it is not clear that an expansion of this type will hold. We seem to get inevitably an extra term of order $O_p(n^{-2/3})$ in (b), which does not fit into this framework.

On the other hand, in our score approach where we use the MLE $\hat{F}_{n,\boldsymbol{\alpha}}$ to estimate F for fixed $\boldsymbol{\alpha}$, our estimator is a kind of hybrid estimator, which requires estimating the argmax MLE $\hat{F}_{n,\boldsymbol{\alpha}}$ of F for fixed $\boldsymbol{\alpha}$ and is defined as the zero of a non-smooth score function as a function of $\boldsymbol{\alpha}$. In the expansion of our score function $\psi_{\epsilon,n}$, we now get

$$\psi_{\epsilon,n}(\hat{\boldsymbol{\alpha}}_n, \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}) = \psi_\epsilon'(\boldsymbol{\alpha}_0, F_0)(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) + \boldsymbol{W}_n + o_p\left(n^{-1/2} + \|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\|\right),$$

where ψ_ϵ' is the matrix representing the total derivative of the population equivalent score function $\boldsymbol{\alpha} \mapsto \psi_\epsilon(\boldsymbol{\alpha}, F_{\boldsymbol{\alpha}})$ (for some deterministic function $F_{\boldsymbol{\alpha}}$ satisfying $F_{\boldsymbol{\alpha}_0} = F_0$) and $\boldsymbol{W}_n$ is a term of order $O_p(n^{-1/2})$. In contrast to the expansion for the argmax statistic Γ_n, extra terms of order $O_p(n^{-2/3})$ do not hurt in the score decomposition since, by definition of the score estimator, we get

$$\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0 \sim -\psi_\epsilon'(\boldsymbol{\alpha}_0, F_0)^{-1}\boldsymbol{W}_n,$$

where $\boldsymbol{Z}_1 \sim \boldsymbol{Z}_2$ is $\boldsymbol{Z}_1 = \boldsymbol{Z}_2 + o_p(1)$. So we have the remarkable situation that finding the estimate $\hat{\boldsymbol{\alpha}}_n$ by a combination of a maximization approach for F and a score approach for $\boldsymbol{\alpha}$, can be proved to give $\sqrt{n}$-consistent estimates of $\boldsymbol{\alpha}_0$, in contrast to the completely argmax approach, using profile likelihood, for which we even still do not know whether it is $\sqrt{n}$-consistent. Despite the fact that the limiting distribution of the MLE of $\boldsymbol{\alpha}_0$ is still unknown, our simulation experiments indicate that, even if the MLE would be $\sqrt{n}$-consistent, its variance is clearly bigger than the other estimates we propose in this chapter.

2.3 Behavior of the maximum likelihood estimator of the distribution function

For fixed $\boldsymbol{\alpha}$, the MLE $\hat{F}_{n,\boldsymbol{\alpha}}$ based on $\ell_n(\boldsymbol{\alpha}, F)$ is a piecewise constant function with jumps at a subset of $\{T_i - \boldsymbol{\alpha}^T\boldsymbol{X}_i : 1 \leq i \leq n\}$. Once we have fixed the parameter

$\boldsymbol{\alpha}$, the order statistics on which the MLE is based are the order statistics of the values $U_1^{\boldsymbol{\alpha}} = T_1 - \boldsymbol{\alpha}^T\boldsymbol{X}_1, \ldots, U_n^{\boldsymbol{\alpha}} = T_n - \boldsymbol{\alpha}^T\boldsymbol{X}_n$ and the values of the corresponding Δ_i denoted by $\Delta_i^{\boldsymbol{\alpha}}$. The MLE can be characterized as the left derivative of the GCM of a cumulative sum diagram consisting of the points $(0,0)$ and

$$\left(i, \sum_{j=1}^{i} \Delta_{(j)}^{\boldsymbol{\alpha}}\right) \qquad 1 \le i \le n,$$

where $\Delta_{(j)}^{\boldsymbol{\alpha}}$ corresponds to the jth order statistic of the $U_i^{\boldsymbol{\alpha}} = T_i - \boldsymbol{\alpha}^T\boldsymbol{X}_i, 1 \le i \le n$ (assuming that no ties are present in the observation times). We have:

$$\mathbb{P}\left\{\Delta_i^{\boldsymbol{\alpha}} = 1 \mid U_i^{\boldsymbol{\alpha}} = u\right\} = \int F_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(x|u)\, d\boldsymbol{x},$$

where $f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\cdot|u)$ denotes the conditional density of $\boldsymbol{X}$ given that $T - \boldsymbol{\alpha}^T\boldsymbol{X} = u$. Hence, defining

$$F_{\boldsymbol{\alpha}}(u) \stackrel{\text{def}}{=} \int F_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|u)\, d\boldsymbol{x}, \tag{2.3.1}$$

we can consider the $\Delta_i^{\boldsymbol{\alpha}}$ as coming from a sample in the ordinary current status model, where the observations are of the form $\left(U_i^{\boldsymbol{\alpha}}, \Delta_i^{\boldsymbol{\alpha}}\right)$, and where the observation times have density $f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}$ and where $\Delta_i^{\boldsymbol{\alpha}} = 1$ with probability $F_{\boldsymbol{\alpha}}(U_i^{\boldsymbol{\alpha}})$ at observation $U_i^{\boldsymbol{\alpha}}$.

Remark 2.3.1. *Assume that T and $\boldsymbol{X}$ are continuous random variables, then we can write*

$$\begin{aligned} F'_{\boldsymbol{\alpha}}(u) = &\int f_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|u)\, d\boldsymbol{x} \\ &+ \int F_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) \frac{\partial}{\partial u} f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|u)\, d\boldsymbol{x}. \end{aligned}$$

Integration by parts on the second term yields:

$$\begin{aligned} &\int F_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) \frac{\partial}{\partial u} f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|u)\, d\boldsymbol{x} \\ &\qquad = -(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T \int f_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) \frac{\partial}{\partial u} F_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|u)\, d\boldsymbol{x}, \end{aligned}$$

where $F_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\cdot|u)$ denotes the conditional distribution function of $\boldsymbol{X}$ given that $T - \boldsymbol{\alpha}^T\boldsymbol{X} = u$. This implies that

$$\begin{aligned} F'_{\boldsymbol{\alpha}}(u) = \int f_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) \Big\{ &f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|u) \\ &- (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T \frac{\partial}{\partial u} F_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|u) \Big\}\, d\boldsymbol{x}. \end{aligned}$$

Assuming that $u \mapsto f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|u)$ *stays away from zero on the support of* f_0, *this implies by a continuity argument that* $F_{\boldsymbol{\alpha}}$ *is monotone increasing on the support of* $F'_{\boldsymbol{\alpha}}$ *for* $\boldsymbol{\alpha}$ *close to* $\boldsymbol{\alpha}_0$.
Also note that we get from the fact that F_0 *is a distribution function with compact support:*

$$\lim_{u\to-\infty} F_{\boldsymbol{\alpha}}(u) = 0 \qquad \textit{and} \qquad \lim_{u\to\infty} F_{\boldsymbol{\alpha}}(u) = 1.$$

So we may assume that $F_{\boldsymbol{\alpha}}$ *is a distribution function for* $\boldsymbol{\alpha}$ *close to* $\boldsymbol{\alpha}_0$. *A similar argument can be used if* $\boldsymbol{X}$ *contains discrete random variables.*

Pictures of the MLE $\hat{F}_{n,\alpha}$, based on the values $T_i - \alpha X_i$, and the corresponding function F_α for the model used in our simulation experiment are shown in Figure 2.1 and compared with F_0. Note that F_α involves both a location shift and a change in shape of F_0.

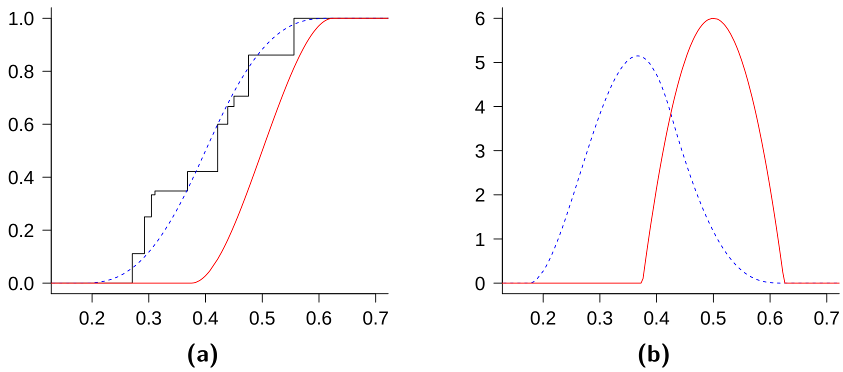


Figure 2.1: (a) The real F_0 (red, solid), the function F_α for $\alpha = 0.6$ (blue, dashed) and the MLE $\hat{F}_{n,\alpha}$ (step function), for a sample of size $n = 1,000$. (b) The real f_0 (red, solid) and the function F'_α for $\alpha = 0.6$ (blue, dashed).

For fixed $\boldsymbol{\alpha}$ in a neighborhood of $\boldsymbol{\alpha}_0$ we can now use standard theory for the MLE from current status theory. The following assumptions are made:

A1. The parameter $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{01}, \dots, \boldsymbol{\alpha}_{0d}) \in \mathbb{R}^d$ is an interior point of Θ and the parameter space Θ is a compact convex set.

A2. $F_{\boldsymbol{\alpha}}$, defined in (2.3.1), has a strictly positive continuous derivative, which stays away from zero on $A_{\epsilon',\boldsymbol{\alpha}} \stackrel{\text{def}}{=} \{u : F_{\boldsymbol{\alpha}}(u) \in [\epsilon', 1-\epsilon']\}$ for all $\boldsymbol{\alpha} \in \Theta$, where $\epsilon' \in (0, \epsilon)$.

A3. The density $u \mapsto f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(u)$ is continuous and also staying away from zero on $A_{\epsilon',\boldsymbol{\alpha}}$ for all $\boldsymbol{\alpha} \in \Theta$, where $A_{\epsilon',\boldsymbol{\alpha}}$ is defined as in A2.

Remark 2.3.2. *Note that the truncation is for the interval $[\epsilon, 1-\epsilon]$, but that we need conditions A2 and A3 to be satisfied for the slightly bigger interval $[\epsilon', 1-\epsilon']$.*

Lemma 2.3.1. *If Assumptions A1, A2 and A3 hold, then:*

(i)

$$\sup_{\boldsymbol{\alpha}\in\Theta} \int \left\{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\}^2 dG(\boldsymbol{x},t) = O_p\left(n^{-2/3}\right).$$

(ii)

$$\mathbb{P}\left(\lim_{n\to\infty} \sup_{\boldsymbol{\alpha}\in\Theta,\, u\in A_{\epsilon',\boldsymbol{\alpha}}} \left|\hat{F}_{n,\boldsymbol{\alpha}}(u) - F_{\boldsymbol{\alpha}}(u)\right| = 0\right) = 1,$$

where ϵ' is chosen as in condition A2 and G is the probability measure of the random variable $(\boldsymbol{X}, T)$.

The proof of Lemma 2.3.1 is given in Appendix A, Section A.1.

2.4 $\sqrt{n}$-consistent regression parameter estimation

2.4.1 A simple estimate based on the MLE $\hat{F}_{n,\boldsymbol{\alpha}}$ without smoothing

We consider the function $\psi_{1\epsilon,n}$ defined by

$$\psi_{1\epsilon,n}(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \int_{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\left\{\delta - \hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\} d\mathbb{P}_n(\boldsymbol{x},t,\delta). \tag{2.4.1}$$

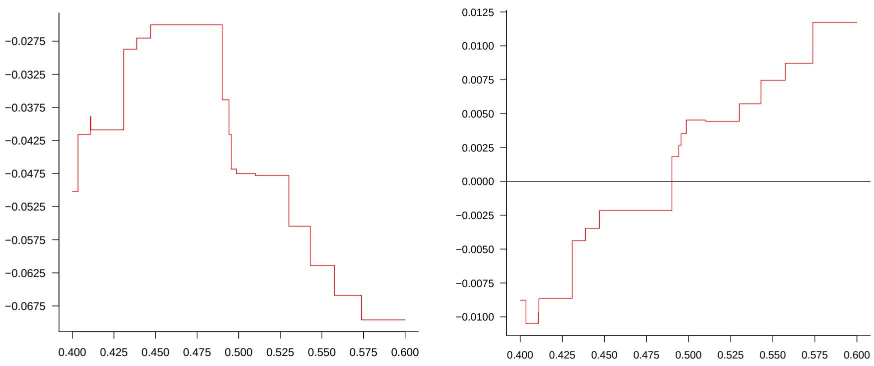
Since the vector $(\hat{F}_{n,\boldsymbol{\alpha}}(T_1-\boldsymbol{\alpha}^T\boldsymbol{X}_1),\dots,\hat{F}_{n,\boldsymbol{\alpha}}(T_n-\boldsymbol{\alpha}^T\boldsymbol{X}_n))^T$ will be the same for all $\boldsymbol{\alpha}$ for which the ranks of the $T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i$ are the same, the function $\psi_{1\epsilon,n}$ can have at most $n!$ different values, for all permutations of the numbers $1,\dots,n$. Figure 2.2b gives a picture of the function $\psi_{1\epsilon,n}$ as a function of α for our simulation model described in Section 2.1.

We would like to define the estimate $\hat{\boldsymbol{\alpha}}_{1n}$ by

$$\psi_{1\epsilon,n}(\hat{\boldsymbol{\alpha}}_{1n}) = \mathbf{0},$$

where $\mathbf{0}$ is the $d-$dimensional vector of zeros, but it is clear that we cannot hope to achieve that due to the discontinuous nature of the score function $\psi_{1\epsilon,n}$. We therefore introduce the following definition.

Definition 2.4.1 (zero-crossing). *We say that $\boldsymbol{\alpha}_*$ is a crossing of zero of a real-valued function $C:\Theta\mapsto\mathbb{R}:\boldsymbol{\alpha}\mapsto C(\boldsymbol{\alpha})$ if each open neighborhood of $\boldsymbol{\alpha}_*$ contains points*



(a) **(b)**

Figure 2.2: The truncated profile log likelihood $\ell_{\epsilon,n}$ (a) and the score function $\psi_{1\epsilon,n}$ (b) as a function of α for a sample of size $n = 100$ and $\epsilon = 0.001$.

$\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \Theta$ such that $\bar{C}(\boldsymbol{\alpha}_1)\bar{C}(\boldsymbol{\alpha}_2) \leq 0$, where $\bar{C}$ is the closure of the image of the function.

Furthermore, we say that a d-dimensional function $\tilde{C} : \Theta \mapsto \mathbb{R}^d : \boldsymbol{\alpha} \mapsto \tilde{C}(\boldsymbol{\alpha}) = (\tilde{C}_1(\boldsymbol{\alpha}), \ldots \tilde{C}_d(\boldsymbol{\alpha}))^T$ has a crossing of zero at a point $\boldsymbol{\alpha}_*$ if $\boldsymbol{\alpha}_*$ is a crossing of zero of each component $\tilde{C}_j : \Theta \mapsto \mathbb{R}, 1 \leq j \leq d$.

Figure 2.2b shows a crossing of zero at a point α close to $\alpha_0 = 0.5$. If the number of dimensions d exceeds one, then a crossing of zero can be thought of as a point $\boldsymbol{\alpha}_* \in \Theta$ such that each component of the score function $\psi_{1\epsilon,n}$ passes through zero in $\boldsymbol{\alpha} = \boldsymbol{\alpha}_*$. Before we state the asymptotic result of our estimator in Theorem 2.4.1, we present in Lemma 2.4.1 below some interesting properties of the population version of the score function.

Lemma 2.4.1. *Let $\psi_{1\epsilon}$ be defined by*

$$\psi_{1\epsilon}(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \int_{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\{\delta - F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\}\, dP_0(\boldsymbol{x}, t, \delta),$$

and define the truncated expectation $\mathbb{E}_{\epsilon,\boldsymbol{\alpha}}$ by

$$\mathbb{E}_{\epsilon,\boldsymbol{\alpha}}(w(\boldsymbol{X}, T, \Delta)) \stackrel{\text{def}}{=} \mathbb{E}\big(1_{\{F_{\boldsymbol{\alpha}}(T-\boldsymbol{\alpha}^T\boldsymbol{X})\in[\epsilon,1-\epsilon]\}} w(\boldsymbol{X}, T, \Delta)\big),$$

for some function w defined on the probability space of the random vector $(\boldsymbol{X}, T, \Delta)$, then:

$$\psi_{1\epsilon}(\boldsymbol{\alpha}_0) = \mathbf{0}$$

and for each $\boldsymbol{\alpha} \in \Theta$ *we have:*

$$(i) \qquad \psi_{1\epsilon}(\boldsymbol{\alpha}) = \mathbb{E}_{\epsilon,\boldsymbol{\alpha}}\left[Cov\left(\Delta, \boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X}\right)\right]$$

$$(ii) \qquad (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\mathbb{E}_{\epsilon,\boldsymbol{\alpha}}\left[Cov\left(\Delta, \boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X}\right)\right] \geq 0 \text{ for all } \boldsymbol{\alpha} \in \Theta,$$

and $\boldsymbol{\alpha}_0$ *is the only value such that (ii) holds. The vector of partial derivatives of* $\psi_{1\epsilon}$ *at* $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ *is given by*

$$\psi'_{1\epsilon}(\boldsymbol{\alpha}_0) = \mathbb{E}_{\epsilon,\boldsymbol{\alpha}_0}\left[f_0(T - \boldsymbol{\alpha}_0^T\boldsymbol{X})Cov(\boldsymbol{X}|T - \boldsymbol{\alpha}_0^T\boldsymbol{X})\right].$$

The proof of Lemma 2.4.1 is given in Appendix A, Section A.2. An illustration of the second result (ii) is given in Figure 2.3, this property is used in the proof of consistency of our estimator $\hat{\boldsymbol{\alpha}}_{1n}$ also provided in Appendix A, Section A.2, as the first part of Theorem 2.4.1.

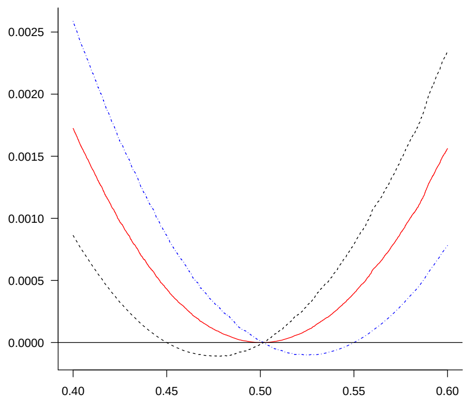


Figure 2.3: The function $\alpha \mapsto (\alpha - \alpha_*)\int_{F_\alpha(t-\alpha x)\in[\epsilon,1-\epsilon]} x\left\{\delta - F_\alpha(t - \alpha x)\right\} dP_0(x, t, \delta)$ as a function of α, with $\alpha_* = 0.45$ (black, dashed), $\alpha_* = \alpha_0 = 0.50$ (red, solid) and $\alpha_* = 0.55$ (blue, dashed-dotted) for $\epsilon = 0.001$.

The following assumptions are also needed for the asymptotic normality results of our estimators:

A4. The function $F_{\boldsymbol{\alpha}}$ is twice continuously differentiable on the interior of the support $S_{\boldsymbol{\alpha}}$ of $f_{\boldsymbol{\alpha}} = F'_{\boldsymbol{\alpha}}$ for all $\boldsymbol{\alpha} \in \Theta$.

A5. The density $f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(u)$ of $T - \boldsymbol{\alpha}^T\boldsymbol{X}$ and the conditional expectations $\mathbb{E}\{\boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X} = u\}$ and $\mathbb{E}\{\boldsymbol{X}\boldsymbol{X}^T|T - \boldsymbol{\alpha}^T\boldsymbol{X} = u\}$ are twice continuously differentiable functions w.r.t. u, except possibly at a finite number of points. The functions $\boldsymbol{\alpha} \mapsto f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(u)$, $\boldsymbol{\alpha} \mapsto \mathbb{E}\{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}=u\}$ and $\boldsymbol{\alpha} \mapsto \mathbb{E}\{\boldsymbol{X}\boldsymbol{X}^T|T-\boldsymbol{\alpha}^T\boldsymbol{X}=u\}$ are continuous functions, for u in the definition domain of the functions and for all $\boldsymbol{\alpha} \in \Theta$. The density of $(\boldsymbol{X}, T)$ has compact support.

Theorem 2.4.1. *Let Assumptions A1-A5 be satisfied and suppose that the covariance* $Cov(\boldsymbol{X}, F_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{X})|T - \boldsymbol{\alpha}^T\boldsymbol{X} = u)$ *is not identically zero for* u *in the region* $A_{\epsilon,\boldsymbol{\alpha}}$*, for each* $\boldsymbol{\alpha} \in \Theta$*. Moreover, let* $\hat{\boldsymbol{\alpha}}_{1n}$ *be defined by a crossing of zero of* $\psi_{1\epsilon,n}$*.*

(i) [Existence of a root] A crossing of zero $\hat{\boldsymbol{\alpha}}_{1n}$ *of* $\psi_{1\epsilon,n}$ *exists with probability tending to one.*

(ii) [Consistency]

$$\hat{\boldsymbol{\alpha}}_{1n} \xrightarrow{p} \boldsymbol{\alpha}_0, \qquad n \to \infty.$$

(iii) [Asymptotic normality] $\sqrt{n}\{\hat{\boldsymbol{\alpha}}_{1n} - \boldsymbol{\alpha}_0\}$ *is asymptotically normal for* $n \to \infty$*, with mean zero and variance* $\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}$*, where*

$$\boldsymbol{A} \stackrel{\text{def}}{=} \mathbb{E}_\epsilon\Big[f_0(T - \boldsymbol{\alpha}_0^T\boldsymbol{X})\, Cov(\boldsymbol{X}|T - \boldsymbol{\alpha}_0^T\boldsymbol{X})\Big],$$

and

$$\boldsymbol{B} \stackrel{\text{def}}{=} \mathbb{E}_\epsilon\Big[F_0(T - \boldsymbol{\alpha}_0^T\boldsymbol{X})\{1 - F_0(T - \boldsymbol{\alpha}_0^T\boldsymbol{X})\}Cov(\boldsymbol{X}|T - \boldsymbol{\alpha}_0^T\boldsymbol{X})\Big],$$

defining $\mathbb{E}_\epsilon(w(\boldsymbol{X}, T, \Delta)) \stackrel{\text{def}}{=} \mathbb{E}\{1_{\{F_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\in[\epsilon,1-\epsilon]\}}w(\boldsymbol{X}, T, \Delta)\}$ *for functions* w *and assuming that* $\boldsymbol{A}$ *is nonsingular.*

Remark 2.4.1. *Note that* $Cov(\boldsymbol{X}, F_0(u+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{X})|T-\boldsymbol{\alpha}^T\boldsymbol{X} = u)$ *is not identically zero for* u *in the region* $\{u : \epsilon \le F_{\boldsymbol{\alpha}}(u) \le 1-\epsilon\}$ *if the conditional distribution of* $\boldsymbol{X}$ *given* $T - \boldsymbol{\alpha}^T\boldsymbol{X} = u$*, is nondegenerate for some* u *in this region if* F_0 *is strictly increasing on* $\{u : \epsilon \le F_{\boldsymbol{\alpha}}(u) \le 1 - \epsilon\}$*.*

Due to the discontinuous nature of the profiled log likelihood ℓ_n and the score function $\psi_{1\epsilon,n}$ (see Figure 2.2), the MLE and the estimator defined in Theorem 2.4.1 are not necessarily unique. The result of Theorem 2.4.1 is valid for any $\hat{\boldsymbol{\alpha}}_{1n}$ satisfying Definition 2.4.1. In the remainder of this thesis we will refer to our estimator defined in Theorem 2.4.1 as the simple score estimator (SSE). In the next section we extend this estimator to an almost efficient estimator of the finite dimensional regression parameter in the current status linear regression model; this estimator will be referred to as the efficient score estimator (ESE).

2.4.2 Efficient estimates involving the MLE $\hat{F}_{n,\boldsymbol{\alpha}}$

Recall from Section 1.1.1 that the efficient score function for Model (2.1) is given by

$$\tilde{\ell}_{\boldsymbol{\alpha}_0,F_0}(\boldsymbol{x},t,\delta) = \left\{\mathbb{E}(\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X} = t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - \boldsymbol{x}\right\} f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) \cdot \left\{\frac{\delta}{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})} - \frac{1-\delta}{1-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})}\right\}.$$

The derivation of the efficient score is given in Section 2.4.5. The idea of the ESE which we will define below, is to obtain an efficient estimate of the regression parameter by constructing an estimate that resembles the root of this efficient score function. Since the function $\tilde{\ell}_{\boldsymbol{\alpha}_0,F_0}$ depends on the density f_0 of the model, we first introduce an estimate of $f_{\boldsymbol{\alpha}} = F_{\boldsymbol{\alpha}}'$ based on the MLE $\hat{F}_{n,\boldsymbol{\alpha}}$. Smoothing is needed to construct our estimate due to the discontinuous nature of the MLE $\hat{F}_{n,\boldsymbol{\alpha}}$.

Let K be a probability density function with derivative K' satisfying

(K1) The probability density K has support [-1,1], is twice continuously differentiable and symmetric on $\mathbb{R}$.

Let $h > 0$ be a smoothing parameter and K_h respectively K_h' be the scaled versions of K and K' respectively, given by

$$K_h(\cdot) = h^{-1}K\left(h^{-1}(\cdot)\right) \quad \text{and} \quad K_h'(\cdot) = h^{-2}K'\left(h^{-1}(\cdot)\right). \tag{2.4.2}$$

Define the density estimate:

$$f_{nh,\boldsymbol{\alpha}}(u) \stackrel{\text{def}}{=} \int K_h(u-w)\,d\hat{F}_{n,\boldsymbol{\alpha}}(w). \tag{2.4.3}$$

We consider

$$\psi_{2\epsilon,nh}(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \int_{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x} f_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) \cdot \frac{\delta - \hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\{1-\hat{F}_{n,\beta}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\}}\,d\mathbb{P}_n(\boldsymbol{x},t,\delta), \tag{2.4.4}$$

and let, analogously to the SSE introduced in the previous section, $\hat{\boldsymbol{\alpha}}_{2n}$ be the estimate of $\boldsymbol{\alpha}_0$, defined by a zero-crossing of the score function $\psi_{2\epsilon,nh}$.

Theorem 2.4.2. *Suppose that the conditions of Theorem 2.4.1 hold and that the function $F_{\boldsymbol{\alpha}}$ is three times continuously differentiable on the interior of the support $S_{\boldsymbol{\alpha}}$. Let $\hat{\boldsymbol{\alpha}}_{2n}$ be defined by a zero-crossing of $\psi_{2\epsilon,nh}$. Then, as $n \to \infty$, and $h \asymp n^{-1/7}$,*

(i) [Existence of a root] A crossing of zero $\hat{\boldsymbol{\alpha}}_{2n}$ of $\psi_{2\epsilon,nh}$ exists with probability tending to one.

(ii) [Consistency]

$$\hat{\boldsymbol{\alpha}}_{2n} \xrightarrow{p} \boldsymbol{\alpha}_0, \qquad n \to \infty.$$

(iii) [Asymptotic normality] $\sqrt{n}\{\hat{\boldsymbol{\alpha}}_{2n} - \boldsymbol{\alpha}_0\}$ *is asymptotically normal for* $n \to \infty$*, with mean zero and variance* $\boldsymbol{I}_\epsilon(\boldsymbol{\alpha}_0)^{-1}$*, where*

$$\boldsymbol{I}_\epsilon(\boldsymbol{\alpha}_0) \stackrel{\text{def}}{=} \mathbb{E}_\epsilon \left\{ \frac{f_0(T - \boldsymbol{\alpha}_0^T \boldsymbol{X})^2\, Cov(\boldsymbol{X}|T - \boldsymbol{\alpha}_0^T \boldsymbol{X})}{F_0(T - \boldsymbol{\alpha}_0^T \boldsymbol{X})\{1 - F_0(T - \boldsymbol{\alpha}_0^T \boldsymbol{X})\}} \right\}, \tag{2.4.5}$$

which is assumed to be nonsingular.

A picture of the score function $\psi_{2\epsilon,nh}$ is shown in Figure 2.4. Note that the range on the vertical axis is considerably larger than the range on the vertical axis of the corresponding score function $\psi_{1\epsilon,n}$.

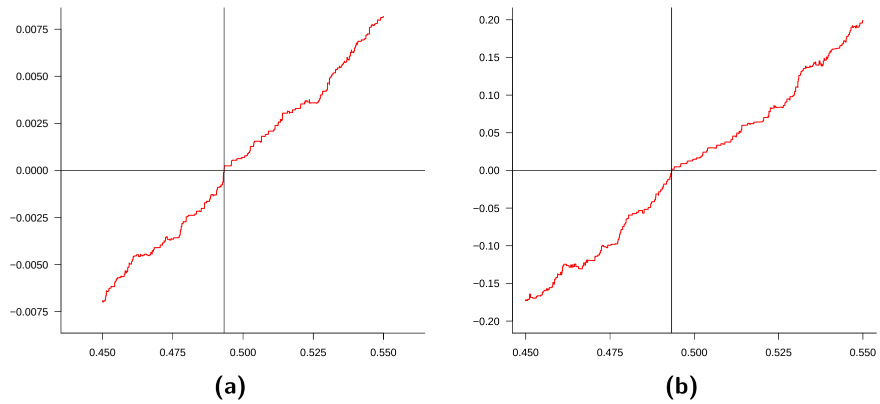


Figure 2.4: The score functions $\psi_{1\epsilon,n}$ (a) and $\psi_{2\epsilon,nh}$ (b) as functions of α for a sample of size $n = 1,000$ with $\epsilon = 0.001$ and $h = 0.5n^{-1/7}$.

2.4.3 Efficient estimates not involving the MLE $\hat{F}_{n,\boldsymbol{\alpha}}$

So far, we defined estimates which depend on the MLE $\hat{F}_{n,\boldsymbol{\alpha}}$. In this section we show that the score approach is also suited for defining efficient estimates that do not depend on the behavior of the MLE $\hat{F}_{n,\boldsymbol{\alpha}}$. If one starts from the log likelihood and first plugs-in a smooth estimate $\tilde{F}_{n,\boldsymbol{\alpha}}$ of F in $\ell_n(\boldsymbol{\alpha}, F)$ then, if $\tilde{F}_{n,\boldsymbol{\alpha}}$ is differentiable w.r.t. $\boldsymbol{\alpha}$, the score can be constructed as the derivative of $\boldsymbol{\alpha} \mapsto \ell_n(\boldsymbol{\alpha}, \tilde{F}_{n,\boldsymbol{\alpha}})$.
Define the plug-in estimate of $F_{\boldsymbol{\alpha}}$ by

$$F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T \boldsymbol{x}) \stackrel{\text{def}}{=} \frac{\int \delta K_h(t - \boldsymbol{\alpha}^T \boldsymbol{x} - u + \boldsymbol{\alpha}^T \boldsymbol{y})\, d\mathbb{P}_n(\boldsymbol{y}, u, \delta)}{\int K_h(t - \boldsymbol{\alpha}^T \boldsymbol{x} - u + \boldsymbol{\alpha}^T \boldsymbol{y})\, d\mathbb{G}_n(\boldsymbol{y}, u)}, \tag{2.4.6}$$

where $\mathbb{G}_n$ is the empirical distribution function of the pairs $(\boldsymbol{X}_i, T_i), 1 \le i \le n$ and where K_h is again a scaled version of a probability density function K, satisfying condition (K1). The plug-in estimates are not necessarily monotone but we show in Theorem 2.4.4 that $F_{nh,\boldsymbol{\alpha}}$ is monotone with probability tending to one as $n \to \infty$ and $\boldsymbol{\alpha} \to \boldsymbol{\alpha}_0$. Another way of writing $F_{nh,\boldsymbol{\alpha}}$ is in terms of ordinary sums. Let

$$g_{nh,1,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \Delta_i K_h(t - \boldsymbol{\alpha}^T\boldsymbol{x} - T_i + \boldsymbol{\alpha}^T\boldsymbol{X}_i),$$

and

$$g_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n K_h(t - \boldsymbol{\alpha}^T\boldsymbol{x} - T_i + \boldsymbol{\alpha}^T\boldsymbol{X}_i),$$

then

$$F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) = \frac{g_{nh,1,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})}{g_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})} = \frac{\sum_{i=1}^n \Delta_i K_h(t - \boldsymbol{\alpha}^T\boldsymbol{x} - T_i + \boldsymbol{\alpha}^T\boldsymbol{X}_i)}{\sum_{i=1}^n K_h(t - \boldsymbol{\alpha}^T\boldsymbol{x} - T_i + \boldsymbol{\alpha}^T\boldsymbol{X}_i)},$$

in which we recognize the Nadaraya-Watson statistic (Nadaraya, 1964 and Watson, 1964). One could also omit the diagonal term $j = i$ in the sums above when estimating $F_{nh,\boldsymbol{\alpha}}(T_i - \boldsymbol{\alpha}^T\boldsymbol{X}_i)$ which is often done in the econometric literature (see e.g. Härdle et al. (1993)). In our computer experiments however, this gave an estimate of the distribution function which had a more irregular behavior than the estimator with the diagonal term included.

If we replace F in the truncated log likelihood $\ell_{\epsilon,n}$ defined in (2.1.3) by $F_{nh,\boldsymbol{\alpha}}$, the truncated log likelihood becomes a function of $\boldsymbol{\alpha}$ only. Although the log likelihood has discontinuities if we consider the lower and upper boundaries $F_{nh,\boldsymbol{\alpha}}^{-1}(\epsilon)$ and $F_{nh,\boldsymbol{\alpha}}^{-1}(1-\epsilon)$ of the integral as a function of $\boldsymbol{\alpha}$, an asymptotic representation of the partial derivatives of the truncated log likelihood is given by the score function

$$\psi_{3\epsilon,nh}(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \int_{F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \frac{\partial_{\boldsymbol{\alpha}} F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\{\delta - F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\}}{F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\{1 - F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\}}\, d\mathbb{P}_n(\boldsymbol{x}, t, \delta), \tag{2.4.7}$$

where the vector of partial derivatives of the plug-in estimate $F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})$, given by (2.4.6), w.r.t. $\boldsymbol{\alpha}$ has the following form:

$$\partial_{\boldsymbol{\alpha}} F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) = \frac{\int (\boldsymbol{y} - \boldsymbol{x})\{\delta - F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\} K_h'(t - \boldsymbol{\alpha}^T\boldsymbol{x} - u + \boldsymbol{\alpha}\boldsymbol{y})\, d\mathbb{P}_n(\boldsymbol{y}, u, \delta)}{g_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})}.$$

From here onwards, we will use the notation $\partial_{\boldsymbol{\alpha}}$ to denote the d-dimensional vector of partial derivatives given by

$$\partial_{\boldsymbol{\alpha}} h(\boldsymbol{\alpha}) = \left(\frac{\partial h}{\partial \alpha_1}(\boldsymbol{\alpha}), \ldots, \frac{\partial h}{\partial \alpha_d}(\boldsymbol{\alpha})\right)^T,$$

for functions $h : \Theta \mapsto \mathbb{R}$. For functions $h : \Theta \mapsto \mathbb{R}^m$, $\partial_{\boldsymbol{\alpha}} h$ denotes the corresponding Jacobian matrix of the map h.

We define the plug-in estimator $\hat{\boldsymbol{\alpha}}_{3n}$ of $\boldsymbol{\alpha}_0$ by

$$\psi_{3\epsilon,nh}(\hat{\boldsymbol{\alpha}}_{3n}) = \mathbf{0}. \tag{2.4.8}$$

A picture of the truncated log likelihood $\alpha \mapsto \ell_{n,\epsilon}(\alpha, F_{nh,\alpha})$ and score function $\psi_{3\epsilon,nh}$ for the plug-in method is shown in Figure 2.5. Since $F_{nh,\alpha}(t - \boldsymbol{\alpha}^T\boldsymbol{x})$ is continuous, we no longer need to introduce the concept of a zero-crossing to ensure existence of the estimator and we can work with the zero of the score function $\psi_{3\epsilon,nh}$ instead. Our main result on the plug-in estimator is given below.

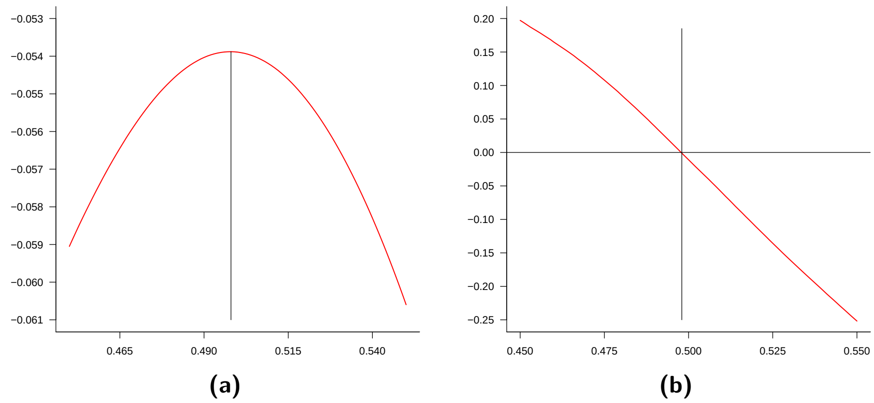


Figure 2.5: The truncated profile log likelihood $l_{\epsilon,n}$ for the plug-in $F_{nh,\alpha}$ (a) and the score function $\psi_{3\epsilon,nh}$ (b) as a function of α for a sample of size $n = 1{,}000$ with $\epsilon = 0.001$ and $h = 0.5n^{-1/5}$.

Theorem 2.4.3. *If Assumptions A1-A5 hold and*

$$-(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T \int_{F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \in [\epsilon, 1-\epsilon]} \frac{\partial_{\boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\{F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x}) - F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\}}{F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\{1 - F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\}} dG(\boldsymbol{x}, t), \tag{2.4.9}$$

is nonzero for each $\boldsymbol{\alpha} \in \Theta$ except for $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$, then we get for $\hat{\boldsymbol{\alpha}}_{3n}$ being the plug-in estimator introduced above, as $n \to \infty$, and $h \asymp n^{-1/5}$:

(i) [Existence of a root] A point $\hat{\boldsymbol{\alpha}}_{3n}$, satisfying (2.4.8), exists with probability tending to one.

(ii) [Consistency]

$$\hat{\boldsymbol{\alpha}}_{3n} \xrightarrow{p} \boldsymbol{\alpha}_0, \qquad n \to \infty.$$

(iii) [Asymptotic normality] $\sqrt{n}\{\hat{\boldsymbol{\alpha}}_{3n} - \boldsymbol{\alpha}_0\}$ is asymptotically normal for $n \to \infty$, with mean zero and variance $\boldsymbol{I}_\epsilon(\boldsymbol{\alpha}_0)^{-1}$ where $\boldsymbol{I}_\epsilon(\boldsymbol{\alpha}_0)$, defined in (2.4.5), is assumed to be nonsingular.

Remark 2.4.2. *Note that using an expansion in $\boldsymbol{\alpha} - \boldsymbol{\alpha}_0$, we can write $\partial_{\boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T \boldsymbol{x})$ as*

$$\begin{aligned}
&\int (\boldsymbol{y} - \boldsymbol{x}) f_0(t - \boldsymbol{\alpha}_0^T \boldsymbol{x} + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T(\boldsymbol{y} - \boldsymbol{x})) f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{y}|t - \boldsymbol{\alpha}^T\boldsymbol{x}) d\boldsymbol{y} \\
&\qquad + \int F_0(t - \boldsymbol{\alpha}_0^T \boldsymbol{x} + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T(\boldsymbol{y} - \boldsymbol{x})) \partial_{\boldsymbol{\alpha}} f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{y}|t - \boldsymbol{\alpha}^T\boldsymbol{x})\, d\boldsymbol{y} \\
&= f_0(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \mathbb{E}\left\{\boldsymbol{X} - \boldsymbol{x} | T - \boldsymbol{\alpha}^T\boldsymbol{X} = t - \boldsymbol{\alpha}^T\boldsymbol{x}\right\} + O(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)
\end{aligned}$$

so that the integral defined in (2.4.9) can be approximated by

$$\begin{aligned}
&-(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T \int_{F_{\boldsymbol{\alpha}}(u) \in [\epsilon, 1-\epsilon]} f_0(u) \mathbb{E}\left\{\boldsymbol{X} - \boldsymbol{x} | T - \boldsymbol{\alpha}^T\boldsymbol{X} = u\right\} \\
&\qquad\qquad \cdot \frac{F_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) - F_{\boldsymbol{\alpha}}(u)}{F_{\boldsymbol{\alpha}}(u)\{1 - F_{\boldsymbol{\alpha}}(u)\}} f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|u)\, d\boldsymbol{x}\, du \\
&= \int_{F_{\boldsymbol{\alpha}}(u) \in [\epsilon, 1-\epsilon]} \frac{f_0(u)\, \mathit{Cov}((\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{X}, F_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{X}) | T - \boldsymbol{\alpha}^T\boldsymbol{X} = u)}{F_{\boldsymbol{\alpha}}(u)\{1 - F_{\boldsymbol{\alpha}}(u)\}}\, du,
\end{aligned}$$

which is positive by the monotonicity of F_0 (see also the discussion in Li and Zhang (1998) about this covariance and the proof of Lemma 2.4.1 given in Appendix A, Section A.2). Figure 2.6 shows the integral in (2.4.9) for our simulation model for $\alpha \in [0.45, 0.55]$ and illustrates that this integral is strictly positive except for $\alpha = \alpha_0 = 0.5$, which is a crucial property for the proof of the consistency of the plug-in estimator given in Appendix A, Section A.4.

We also have the following results for the plug-in estimator.

Theorem 2.4.4. *Let the conditions of Theorem 2.4.3 be satisfied, then we have on each interval I contained in the support of $f_{\boldsymbol{\alpha}}$ and for each $\boldsymbol{\alpha} \in \Theta$:*

$$P\left\{F_{nh,\boldsymbol{\alpha}} \text{ is monotonically increasing on } I\right\} \xrightarrow{p} 1.$$

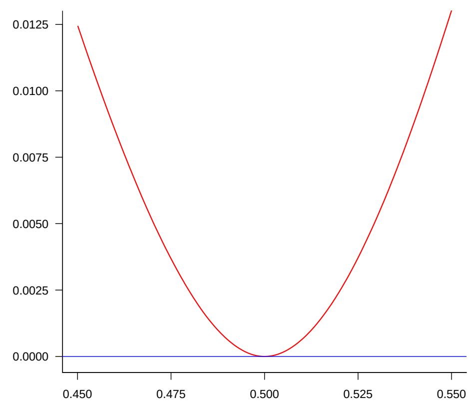


Figure 2.6: The integral defined in (2.4.9), as a function of α, with $\epsilon = 0.001$.

The proof of Theorem 2.4.4 follows from the asymptotic monotonicity of the plug-in estimate in the classical current status model (without regression parameters) and is proved in the same way as Theorem 3.3 of Groeneboom et al. (2010).

Theorem 2.4.5. *Let the conditions of Theorem 2.4.3 be satisfied. Then, for $\hat{\alpha}_{3n}$ being the plug-in estimator of $\boldsymbol{\alpha}_0$,*

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_{3n} - \boldsymbol{\alpha}_0) = n^{-1/2} \boldsymbol{I}_\epsilon(\boldsymbol{\alpha}_0)^{-1} \sum_{i \in J_{F_0}} f_0(T_i - \boldsymbol{\alpha}_0^T \boldsymbol{X}_i)\{\mathbb{E}(\boldsymbol{X}_i | T_i - \boldsymbol{\alpha}_0^T \boldsymbol{X}_i) - \boldsymbol{X}_i\} \cdot \frac{\Delta_i - F_0(T_i - \boldsymbol{\alpha}_0^T \boldsymbol{X}_i)}{F_0(T_i - \boldsymbol{\alpha}_0^T \boldsymbol{X}_i)\{1 - F_0(T_i - \boldsymbol{\alpha}_0^T \boldsymbol{X}_i)\}} + o_p(1).$$

where $J_H = \{i : \epsilon \le H(T_i - \boldsymbol{\alpha}_0^T \boldsymbol{X}_i) \le 1 - \epsilon\}$ for some function H.

The representation of Theorem 2.4.5 plays an important role in determining the variance of smooth functionals, of which the intercept $\int u \, dF_0(u)$ is an example. The proof of Theorem 2.4.5 is given in Appendix A, Section A.4.1. A similar representation holds for the estimators defined in Theorem 2.4.1 and Theorem 2.4.2 (see the proofs of Theorem 2.4.1 and 2.4.2 given in Appendix A, Sections A.2 and A.3 respectively).

Remark 2.4.3. *The plug-in method also suggests the use of U-statistics. By straightfor-*

ward calculations, we can write the score function defined in (2.4.7) as

$$\begin{aligned}
\psi_{3\epsilon,nh}(\boldsymbol{\alpha}) &= \frac{1}{n^2}\sum_{i\in J_{F_{nh,\boldsymbol{\alpha}}}} \frac{\partial_{\boldsymbol{\alpha}} F_{nh,\boldsymbol{\alpha}}(T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i)\{\Delta_i - F_{nh,\boldsymbol{\alpha}}(T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i)\}}{F_{nh,\boldsymbol{\alpha}}(T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i)\{1-F_{nh,\boldsymbol{\alpha}}(T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i)\}} \\
&= \frac{1}{n^2}\sum_{i\in J_{F_{nh,\boldsymbol{\alpha}}}}\sum_{j\neq i} \frac{\Delta_i\Delta_j(\boldsymbol{X}_j-\boldsymbol{X}_i)K_h'(T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i-T_j+\boldsymbol{\alpha}^T\boldsymbol{X}_j)}{g_{nh,1,\boldsymbol{\alpha}}(T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i)} \\
&\quad + \frac{1}{n^2}\sum_{i\in J_{F_{nh,\boldsymbol{\alpha}}}}\sum_{j\neq i} \frac{(1-\Delta_i)(1-\Delta_j)(\boldsymbol{X}_j-\boldsymbol{X}_i)K_h'(T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i-T_j+\boldsymbol{\alpha}^T\boldsymbol{X}_j)}{g_{nh,0,\boldsymbol{\alpha}}(T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i)} \\
&\quad - \frac{1}{n^2}\sum_{i\in J_{F_{nh,\boldsymbol{\alpha}}}}\sum_{j\neq i} \frac{(\boldsymbol{X}_j-\boldsymbol{X}_i)K_h'(T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i-T_j+\boldsymbol{\alpha}^T\boldsymbol{X}_j)}{g_{nh,\boldsymbol{\alpha}}(T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i)}, \qquad (2.4.10)
\end{aligned}$$

where $g_{nh,0,\boldsymbol{\alpha}} = g_{nh,\boldsymbol{\alpha}} - g_{nh,1,\boldsymbol{\alpha}}$. Each of the three terms on the right-hand side of (2.4.10) can be rewritten in terms of a scaled second order U-statistic. A proof based on U-statistics requires lengthy and tedious calculations which are avoided in the current approach for proving Theorem 2.4.3. The representation given in Theorem 2.4.5 also indicates that the U-statistics representation does not give the most natural approach to the proof of asymptotic normality and efficiency of $\hat{\boldsymbol{\alpha}}_{3n}$. For these reasons, we do not further examine the results on U-statistics.

Remark 2.4.4. *We propose the bandwidths $h \asymp n^{-1/7}$ respectively $h \asymp n^{-1/5}$ in Theorem 2.4.2 respectively Theorem 2.4.3, which are the usual bandwidths with ordinary second order kernels for the estimates of a density respectively distribution function. Unfortunately, various advices are given in the literature on what smoothing parameters should be used. Klein and Spady (1993) have fourth order kernels and use bandwidths between the orders $n^{-1/6}$ and $n^{-1/8}$ for the estimation of F. Note that the use of fourth order kernels needs the associated functions to have four derivatives in order to have the desired bias reduction. Cosslett (2007) advises a bandwidth h such that $n^{-1/5} \ll h \ll n^{-1/8}$. Both ranges are considerably large and exclude our bandwidth choice $h \asymp n^{-1/5}$. Murphy et al. (1999) considers a penalized maximum likelihood estimator where the penalty parameter λ_n satisfies $1/\lambda_n = O_p\left(n^{2/5}\right)$ and $\lambda_n^2 = o_p\left(n^{-1/2}\right)$. Translating this into bandwidth choice (using $h_n \asymp \sqrt{\lambda_n}$), the conditions correspond to: $n^{-1/5} \lesssim h \ll n^{-1/8}$, suggesting that their conditions do allow the choice $h \asymp n^{-1/5}$ for estimating the distribution function.*

2.4.4 Truncation

We introduce a truncation device in order to avoid unbounded score functions and numerical difficulties. If one starts with the efficient score equation or an estimate thereof, the solution sometimes suggested in the literature is to add a constant c_n, tending to zero as $n \to \infty$, to the factor $F(t - \boldsymbol{\alpha}^T)\{1 - F(t - \boldsymbol{\alpha}^T \boldsymbol{x})\}$ which inevitably will appear in the denominator. This is done in, e.g. Li and Zhang (1998); similar ideas involving a sequence (c_n) are used in Klein and Spady (1993) and Cosslett (2007).

In contrast with the usual approaches to deal with truncation, which imply the selection of a suitable sequence c_n, we do not consider a vanishing truncation sequence but work with a subsample of the data depending on the ϵ and $(1 - \epsilon)$ quantiles of the distribution function estimate for small but fixed $\epsilon \in (0, 1/2)$. This simple device in (2.1.3) moreover implies keeping the characterizing properties of the MLE (see Proposition 1.1 on p. 39 of Groeneboom and Wellner (1992)) which are lost when a vanishing sequence is considered. It is perhaps somewhat remarkable that we can, instead of letting $\epsilon \downarrow 0$, fix $\epsilon > 0$ and still have consistency of our estimators; on the other hand, the estimate proposed by Murphy et al. (1999) is also identified via a subset of the support of the distribution F_0.
Although the truncation area depends on $\boldsymbol{\alpha}$, we show in Appendix A, Section A.2 (see the proof of Theorem 2.4.1) that the population version of the score function, given by

$$\psi_\epsilon(\boldsymbol{\alpha}) = \int_{F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T \boldsymbol{x}) \in [\epsilon, 1-\epsilon]} \phi(\boldsymbol{x}, t, \delta)\{\delta - F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T \boldsymbol{x})\} dP_0(\boldsymbol{x}, t, \delta), \tag{2.4.11}$$

has a derivative at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ that only involves the derivative of the integrand in (2.4.11), but does not involve terms arising from the truncation limits appearing in the integral. Using the truncation in the maximum log likelihood approach would not lead to a derivative of the population version of the log likelihood which ignores the boundaries and therefore this truncation is less suited for argmax estimators.

A drawback of our fixed truncation parameter approach is that we get truncated information. The resulting estimates are therefore not efficient in the classical sense of efficiency but the difference between the efficient variance and almost (determined by the size of ϵ) efficient variance is rather small in our simulation models. A derivation of the efficient information for the semiparametric current status linear regression model is given in Section 2.4.5. We also tried to program the fully efficient estimators proposed by Li and Zhang (1998) and compared their performance to the performance of our almost-efficient estimators. The comparison showed that our estimates perform better in finite samples. Moreover, the estimates by Li and Zhang (1998) involve several kernel density estimates (based on 5 double summations over the data points), resulting in a very large computation time compared to our simple estimates.

The usual conditions in the theory of estimation of F_0 under current status and, more generally, interval censored data are that F_0 corresponds to a distribution with compact support. Otherwise, certain variances easily get infinite, and similarly, the Fisher information in our model can easily become infinite. Truncating by keeping the quantiles between ϵ and $1-\epsilon$ avoids difficulties in this case and allows us to apply the theory which presently has been developed for the current status model.

Note that the score function defined in (2.4.1) does not contain a factor $F(t-\boldsymbol{\alpha}^T\boldsymbol{x})$ or $1-F(t-\boldsymbol{\alpha}^T\boldsymbol{x})$ in the denominator. For simplicity of the proofs, we still impose the truncation area, since the classical results for the current status model are derived under the assumption that the density f_0 is bounded away from zero . We conjecture however that the result of Theorem 2.4.1 remains valid when taking $\epsilon = 0$.

2.4.5 Efficient information in the current status linear regression model

In this section we give the efficiency calculations for the current status linear regression model. The density of one observation in the current status linear regression model is

$$p_{\boldsymbol{\alpha},F}(\boldsymbol{x},t,\delta) = F\left(t-\boldsymbol{\alpha}^T\boldsymbol{x}\right)^{\delta}\left\{1-F\left(t-\boldsymbol{\alpha}^T\boldsymbol{x}\right)\right\}^{1-\delta} f_{\boldsymbol{X},T}(\boldsymbol{x},t).$$

We assume that the distribution of $(\boldsymbol{X},T)$ does not depend on $(\boldsymbol{\alpha},F)$ which implies that the relevant part of the log likelihood is given by

$$l_n(\boldsymbol{\alpha},F) = \sum_{i=1}^{n}\left[\Delta_i \log F(T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i) + (1-\Delta_i)\log\{1-F(T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i)\}\right].$$

If the distribution F is known (parametric case), the information $\boldsymbol{I}_P$ for $\boldsymbol{\alpha}$ is given by

$$\boldsymbol{I}_P(\boldsymbol{\alpha}) = \mathbb{E}\left(\left(\partial_{\boldsymbol{\alpha}}\log p_{\boldsymbol{\alpha},F}(\boldsymbol{X},T,\Delta)\right)\left(\partial_{\boldsymbol{\alpha}}\log p_{\boldsymbol{\alpha},F}(\boldsymbol{X},T,\Delta)\right)^T\right).$$

Straightforward calculations yield that

$$\boldsymbol{I}_P(\boldsymbol{\alpha})_{ij} = \int \frac{\mathbb{E}(X_iX_j|T-\boldsymbol{\alpha}^T\boldsymbol{X}=u)}{F(u)\{1-F(u)\}} f(u)^2 f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(u)\,du,$$

where $f = F'$ and where $f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}$ is the density of $T-\boldsymbol{\alpha}^T\boldsymbol{X}$. When F is unknown, we need to calculate the efficient score function. Let F and P_0 be the probability measures of ε and $(\boldsymbol{X},T,\Delta)$ respectively and let $L_2^0(Q)$ be the Hilbert space of square integrable functions a with respect to the measure dQ satisfying $\int a dQ = 0$. The score operator $l_F : L_2^0(F) \mapsto L_2^0(P_0)$ is defined by

$$\begin{aligned}[l_F a](\boldsymbol{x},t,\delta) &= \mathbb{E}\big(a(\varepsilon)|(\boldsymbol{X},T,\Delta)=(\boldsymbol{x},t,\delta)\big)\\ &= \frac{\delta\int_{-\infty}^{t-\boldsymbol{\alpha}^T\boldsymbol{x}} a(s)dF(s)}{F(t-\boldsymbol{\alpha}^T\boldsymbol{x})} - \frac{(1-\delta)\int_{-\infty}^{t-\boldsymbol{\alpha}^T\boldsymbol{x}} a(s)dF(s)}{1-F(t-\boldsymbol{\alpha}^T\boldsymbol{x})},\end{aligned}$$

with adjoint,

$$[l_F^* b](e) = \mathbb{E}\left(b(\boldsymbol{X}, T, \Delta) | \varepsilon = e\right).$$

The information $\boldsymbol{I}$ for $\boldsymbol{\alpha}$ in the semi-parametric model is defined by

$$\boldsymbol{I}(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \mathbb{E}\left(\tilde{\ell}_{\boldsymbol{\alpha},F}(\boldsymbol{X}, T, \Delta)\tilde{\ell}_{\boldsymbol{\alpha},F}(\boldsymbol{X}, T, \Delta)^T\right),$$

where $\tilde{\ell}_{\boldsymbol{\alpha},F}(\boldsymbol{x}, t, \delta)$ is the efficient score function defined by

$$\tilde{\ell}_{\boldsymbol{\alpha},F}(\boldsymbol{x}, t, \delta) = \ell_{\boldsymbol{\alpha}}(\boldsymbol{x}, t, \delta) - [\ell_F a_*](\boldsymbol{x}, t, \delta),$$

where

$$\ell_{\boldsymbol{\alpha}}(\boldsymbol{x}, t, \delta) = \partial_{\boldsymbol{\alpha}} \log p_{\boldsymbol{\alpha},F}(\boldsymbol{x}, t, \delta) = \frac{-\delta \boldsymbol{x} f(t - \boldsymbol{\alpha}^T \boldsymbol{x})}{F(t - \boldsymbol{\alpha}^T \boldsymbol{x})} + \frac{(1-\delta)\boldsymbol{x} f(t - \boldsymbol{\alpha}^T \boldsymbol{x})}{1 - F(t - \boldsymbol{\alpha}^T \boldsymbol{x})},$$

and $\ell_F a_*$ satisfies

$$\ell_F^* \ell_F a_* = \ell_F^* \ell_{\boldsymbol{\alpha}}. \tag{2.4.12}$$

The efficient score $\tilde{\ell}_{\boldsymbol{\alpha},F}$ can be interpreted as the residual of $\ell_{\boldsymbol{\alpha}}$ projected in the space spanned by $\ell_F a$ for $a \in L_2^0(F)$. Note that, as a consequence of (2.4.12), the efficient information equals

$$\boldsymbol{I}(\boldsymbol{\alpha}) = \mathbb{E}\left(\tilde{\ell}_{\boldsymbol{\alpha},F}(\boldsymbol{X}, T, \Delta)\ell_{\boldsymbol{\alpha}}(\boldsymbol{X}, T, \Delta)^T\right).$$

To find a_*, we have to solve (2.4.12):

$$\begin{aligned}
\ell_F^* \ell_F a_*(e) &= \int_e^\infty \frac{\phi(u)}{F(u)} f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(u)\,du - \int_{-\infty}^e \frac{\phi(u)}{1-F(u)} f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(u)\,du \\
&= -\int_e^{+\infty} \frac{\mathbb{E}(\boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X} = u) f(u)}{1-F(u)} f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(u)\,du \\
&\quad + \int_{-\infty}^e \frac{\mathbb{E}(\boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X} = u) f(u)}{1-F(u)} f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(u)\,du \\
&= \ell_F^* \ell_{\boldsymbol{\alpha}}(e),
\end{aligned} \tag{2.4.13}$$

where $\phi(t) = \int_{-\infty}^t a(s)dF(s)$. Equation (2.4.13) is satisfied with

$$\phi(u) = -\mathbb{E}(\boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X} = u) f(u).$$

Any a_* that satisfies the above equation satisfies (2.4.12) and we get

$$\begin{aligned}
\tilde{\ell}_{\boldsymbol{\alpha},F}(\boldsymbol{x}, t, \delta) = &\left\{\mathbb{E}(\boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X} = t - \boldsymbol{\alpha}^T\boldsymbol{x}) - \boldsymbol{x}\right\} f(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \\
&\cdot \left\{\frac{\delta}{F(t - \boldsymbol{\alpha}^T\boldsymbol{x})} - \frac{1-\delta}{1 - F(t - \boldsymbol{\alpha}^T\boldsymbol{x})}\right\},
\end{aligned}$$

and

$$\boldsymbol{I}(\boldsymbol{\alpha})_{ij} = \int \frac{\mathrm{Cov}(X_i, X_j | T - \boldsymbol{\alpha}^T \boldsymbol{X} = u)}{F(u)\{1 - F(u)\}} f(u)^2 f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(u)\, du. \tag{2.4.14}$$

Note that $\boldsymbol{I}(\boldsymbol{\alpha})^{-1} - \boldsymbol{I}_P(\boldsymbol{\alpha})^{-1}$ equals the minimal increase of the variance of an estimator of $\boldsymbol{\alpha}$ based on an unknown F (semi-parametric case) compared to the situation where F is known (parametric). In our simulation example $\boldsymbol{I}_P(\alpha) = 26.3667$ and $\boldsymbol{I}(\alpha) = 6.5917$.

2.5 Estimation of the intercept

We want to estimate the intercept in model (2.1.1) defined by

$$\mu_0 = \int u\, dF_0(u). \tag{2.5.1}$$

We can replace F_0 in the expression above by the plug-in estimate $F_{nh,\hat{\boldsymbol{\alpha}}_{3n}}$ where $\hat{\boldsymbol{\alpha}}_{3n}$ is the plug-in estimator of $\boldsymbol{\alpha}_0$ defined in (2.4.8). However, to avoid bias in estimating μ_0, we have to estimate F_0 with a smaller bandwidth h, satisfying $h \ll n^{-1/4}$, for example $h \asymp n^{-1/3}$. The matter is discussed in Cosslett (2007), p. 1253.
We have the following result of which the proof can be found in Appendix A, Section A.4.2.

Theorem 2.5.1. *Let the conditions of Theorem 2.4.3 be satisfied, and let $\hat{\boldsymbol{\alpha}}_{3n}$ be the d-dimensional estimate of $\boldsymbol{\alpha}_0$ as obtained by the score procedure, described in Theorem 2.4.3, using a bandwidth of order $n^{-1/5}$. Let $F_{nh,\hat{\boldsymbol{\alpha}}_{3n}}$ be a plug-in estimate of F_0, using $\hat{\boldsymbol{\alpha}}_{3n}$ as the estimate of $\boldsymbol{\alpha}_0$, but using a bandwidth h of order $n^{-1/3}$ instead of $n^{-1/5}$. Finally, let $\hat{\mu}_n$ be the estimate of μ_0, defined by*

$$\hat{\mu}_n \stackrel{\mathrm{def}}{=} \int u\, dF_{nh,\hat{\boldsymbol{\alpha}}_{3n}}(u).$$

Then $\sqrt{n}(\hat{\mu}_n - \mu_0)$ is asymptotically normal, with expectation zero and variance

$$\sigma^2 = a(\boldsymbol{\alpha}_0)' \boldsymbol{I}_\epsilon(\boldsymbol{\alpha}_0)^{-1}\, a(\boldsymbol{\alpha}_0) + \int \frac{F_0(v)\{1 - F_0(v)\}}{f_{T-\boldsymbol{\alpha}_0^T\boldsymbol{X}}(v)}\, dv, \tag{2.5.2}$$

where $a(\boldsymbol{\alpha}_0)$ is the d-dimensional vector, defined by

$$a(\boldsymbol{\alpha}_0) \stackrel{\mathrm{def}}{=} \int \mathbb{E}\{\boldsymbol{X} | T - \boldsymbol{\alpha}_0^T \boldsymbol{X} = u\} f_0(u)\, du,$$

and $\boldsymbol{I}_\epsilon(\boldsymbol{\alpha}_0)$ is defined in (2.4.5).

Remark 2.5.1. *We choose the bandwidth of order $n^{-1/3}$, but other choices are also possible. We can in fact choose $n^{-1/2} \ll h \ll n^{-1/4}$ (see the proof of Theorem 2.5.1 in Appendix A, Section A.4.2). The bandwidth of order $n^{-1/3}$ corresponds to the automatic bandwidth choice of the MLE of F_0, also using the estimate $\hat{\boldsymbol{\alpha}}_{3n}$ of $\boldsymbol{\alpha}_0$.*

Remark 2.5.2. *Note that the variance corresponds to the information lower bound for smooth functionals in the binary choice model, given in Cosslett (2007). The second part of the expression for the variance on the right-hand side of (2.5.2) is familiar from current status theory, see e.g. (10.7), p. 287 of Groeneboom and Jongbloed (2014).*

Instead of considering the plug-in estimate, we could also consider the SSE or ESE. After having determined an estimate $\hat{\boldsymbol{\alpha}}_n$ in this way, we next estimate μ_0 by

$$\hat{\mu}_n = \int u \, d\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u), \tag{2.5.3}$$

where $\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}$ is the MLE corresponding to the estimate $\hat{\boldsymbol{\alpha}}_n$. The theoretical justification of this approach can be proven using the asymptotic theory of smooth functionals given in Groeneboom and Jongbloed (2014), p.286. Using the MLE $\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}$ instead of the plug-in $F_{nh,\hat{\boldsymbol{\alpha}}_n}$ as an estimate of the distribution function F_0, avoids the selection of a bandwidth parameter for the intercept estimate. We discuss in the next section how the bandwidth can be selected by the practitioner in a real data sample.

2.6 Computation and simulations

The computation of our estimates is relatively straightforward in all cases. For the SSE and the ESE, we first compute the MLE for fixed $\boldsymbol{\alpha}$ by the so-called "pool adjacent violators" algorithm of Ayer et al. (1955) for computing the convex minorant of the corresponding cumulative sum diagram. If the MLE has been computed for fixed $\boldsymbol{\alpha}$, we can compute the density estimate f_{nh}. The estimate of $\boldsymbol{\alpha}_0$ is then determined by a root-finding algorithm such as Brent's method. Computation is very fast. For the plug-in estimate, we simply compute the estimate $F_{nh,\boldsymbol{\alpha}}$ as a ratio of two kernel estimators for fixed $\boldsymbol{\alpha}$ and then compute the derivative w.r.t. $\boldsymbol{\alpha}$. Next we use again a root-finding algorithm to determine the zero of the corresponding score function.

Some results from the simulations of our model are available in Table 2.1, which contains the mean value of the estimate, averaged over $N = 10,000$ iterations, and n times the variance of the estimate of $\alpha_0 = 0.5$ (respectively $\mu_0 = 0.5$) for the different methods described above, as well as for the classical MLE of α_0, for different sample sizes n and a truncation parameter $\epsilon = 0.001$. We took the bandwidth $h = 0.5n^{-1/7}$ for the ESE in Section 2.4.1. The bandwidth $h = 0.5n^{-1/5}$ for the plug-in estimate of Section 2.4.3 was chosen based on an investigation of the mean squared error (MSE) for different choices of c in $h = cn^{-1/5}$. Details on how to choose the bandwidth in practice are given in Section 2.7. The true asymptotic values for the variance of $\sqrt{n}(\hat{\alpha}_n - \alpha_0)$ in our simulation model, obtained via the inverse of the information $\boldsymbol{I}_\epsilon(\alpha_0)$, are 0.151707 without truncation and 0.158699 for $\epsilon = 0.001$ and 0.17596 for $\epsilon = 0.01$. We advise to use a truncation parameter ϵ of 0.001 or smaller in practice. The variance defined in Theorem 2.4.1 for $\epsilon = 0.001$ is 0.193612. The lower bounds for the variance of the intercept are 0.257898 for the simple score method and 0.222984 for the efficient methods. Our results show convergence to these bounds.

Table 2.1 shows that the ESE and the plug-in estimates perform reasonably well. A drawback of the plug-in method however is the long computing time for large sample sizes, whereas the computation for the MLE-based estimates is fast even for the larger samples. Note moreover that the plug-in estimate of the distribution function is only asymptotically monotone whereas the MLE is monotone by definition. All our proposed estimates perform better than the classical MLE. The log likelihood for the MLE has a rough behavior, with a larger chance that optimization algorithms might calculate a local maximizer instead of the global one.

The performance of the score estimates is worse than the performance of the plug-in estimates for small sample sizes but increases considerably when the sample size increases. Although the asymptotic variance of the SSE is larger than the almost (determined by the truncation parameter ϵ) efficient variance, the results obtained with this method are noteworthy seen the fact that no smoothing is involved in this simple estimation technique.

Table 2.1 does not provide strong evidence of/against the $\sqrt{n}$-consistency of the classical MLE. Considering the drawbacks of the classical MLE, we advise the use of the plug-in estimate for small sample sizes and the use of the score estimates, based on the MLE, for larger sample sizes, for estimating the parameter $\boldsymbol{\alpha}_0$. We finally suggest to estimate the parameter μ_0 via the MLE corresponding to this $\boldsymbol{\alpha}_0$ estimate, avoiding in this way the bias problem for the kernel estimates of μ_0.

2.7 Bandwidth selection

In this section we discuss the bandwidth selection for the plug-in estimate. A similar idea can be used for the selection of the bandwidth used for the second estimate defined in Section 2.4.2. We define the optimal constant c_{opt} in $h = cn^{-1/5}$ as the minimizer of the mean squared error (MSE),

$$c_{opt} = \arg\min_c MSE(c) \stackrel{\text{def}}{=} \arg\min_c \mathbb{E}\|\hat{\boldsymbol{\alpha}}_{n,h_c} - \boldsymbol{\alpha}_0\|^2,$$

Table 2.1: The mean value and n times the variance of the estimates of α_0 and μ_0 for the simple score estimator (SSE), the efficient score estimator (ESE), the plug-in estimator (plug-in) and the maximum likelihood estimator (MLE) for different sample sizes n. $h_\alpha = 0.5n^{-1/7}$ (for the ESE) and $h_\alpha = 0.5n^{-1/5}$, $h_\mu = 0.75n^{-1/3}$ (for the plug-in), $\epsilon = 0.001$ and $N = 10,000$. The line, preceded by ∞, gives the asymptotic values.

		SSE		ESE		Plug-in		MLE	
	n	mean	$n\times$var	mean	$n\times$var	mean	$n\times$var	mean	$n\times$var
α	100	0.500212	0.364558	0.502247	0.410449	0.499562	0.245172	0.489690	0.307961
	500	0.499845	0.221484	0.499825	0.230178	0.498857	0.191857	0.499315	0.228335
	1,000	0.499982	0.211608	0.500353	0.208102	0.499502	0.192223	0.499937	0.228420
	5,000	0.499901	0.195294	0.499964	0.184807	0.500314	0.181421	0.499933	0.239898
	10,000	0.499988	0.191115	0.499985	0.172758	0.500120	0.172043	0.499994	0.227222
	20,000	0.500038	0.187616	0.500023	0.169762	0.500096	0.174197	0.499952	0.238400
	∞	0.500000	0.193612	0.500000	0.158699	0.500000	0.158699	0.500000	?
μ	100	0.511937	0.468415	0.509679	0.515638	0.495709	0.332949	0.523103	0.425614
	500	0.502258	0.293585	0.502506	0.287576	0.498932	0.254040	0.502514	0.304540
	1,000	0.500839	0.284958	0.500616	0.262684	0.498385	0.270085	0.500937	0.300201
	5,000	0.500345	0.262566	0.500316	0.244892	0.501597	0.241294	0.500270	0.303754
	10,000	0.500127	0.256983	0.500134	0.232973	0.501680	0.245993	0.500076	0.289905
	20,000	0.500020	0.250720	0.500042	0.230901	0.501660	0.244042	0.500101	0.302824
	∞	0.500000	0.257898	0.500000	0.222984	0.500000	0.222984	0.500000	?

where $\hat{\boldsymbol{\alpha}}_{n,h_c}$ is the estimate $\hat{\boldsymbol{\alpha}}_{3n}$ obtained when the constant c in $h = cn^{-1/5}$ is chosen in the estimation method. A picture of the Monte Carlo estimate of MSE as a function of c is shown for the plug-in method in Figure 2.7, where we estimated MSE(c) on a grid $c =$ 0.01, 0.05, 0.10, $\cdots$, 0.95, for a sample size $n = 1,000$ and truncation parameter $\epsilon = 0.001$ by a Monte Carlo experiment with $N = 1,000$ simulation runs,

$$\widehat{MSE(c)} = N^{-1} \sum_{j=1}^{N} \|\hat{\boldsymbol{\alpha}}_{n,h_c}^j - \boldsymbol{\alpha}_0\|^2, \tag{2.7.1}$$

where $\hat{\boldsymbol{\alpha}}_{n,h_c}^j$ is the estimate of $\boldsymbol{\alpha}_0$ in the j-th simulation run, $1 \le j \le, N$.

Since F_0 and $\boldsymbol{\alpha}_0$ are unknown in practice, we cannot compute the actual MSE. We use the bootstrap method proposed by Sen and Xu (2015) to obtain an estimate of MSE. Our proposed estimate $F_{nh,\boldsymbol{\alpha}}$ of the distribution function F_0 satisfies the conditions of

Theorem 3 in Sen and Xu (2015) and the consistency of the bootstrap is guaranteed. Note that it follows from Kosorok (2008) and Sen et al. (2010) that naive bootstrapping, by resampling with replacement $(\boldsymbol{X}_i, T_i, \Delta_i)$, or by generating bootstrap samples from the MLE, is inconsistent for reproducing the distribution of the MLE.
The method works as follows. We let $h_0 = c_0 n^{-1/5}$ be an initial choice of the bandwidth and calculate the plug-in estimates $\hat{\boldsymbol{\alpha}}_{n,h_0}$ and $F_{n,h_{c_0}}$ based on the original sample $(\boldsymbol{X}_i, T_i, \Delta_i), 1 \le i \le n$. We generate a bootstrap sample $(\boldsymbol{X}_i, T_i, \Delta_i^*), 1 \le i \le n$ where the $(\boldsymbol{X}_i, T_i)$ correspond to the $(\boldsymbol{X}_i, T_i)$ in the original sample and where the indicator Δ_i^* is generated from a Bernoulli distribution with probability $F_{n,h_0}(T_i - \hat{\boldsymbol{\alpha}}_{n,h_0}^T \boldsymbol{X}_i)$, and next estimate $\hat{\boldsymbol{\alpha}}_{n,h_c}^*$ from this bootstrap sample. We repeat this B times and estimate MSE(c) by

$$\widehat{MSE}_B(c) = B^{-1} \sum_{b=1}^{B} \|\hat{\boldsymbol{\alpha}}_{n,h_c}^{*b} - \hat{\boldsymbol{\alpha}}_{n,h_{c_0}}\|^2, \tag{2.7.2}$$

where $\hat{\boldsymbol{\alpha}}_{n,h_c}^{*b}$ is the bootstrap estimate in the b-th bootstrap run. The optimal bandwidth $\hat{h}_{opt} = \hat{c}_{opt} n^{-1/5}$ where $\hat{c}_{opt}$ is defined as the minimizer of $\widehat{MSE}_B(c)$.

We analyze the behavior of the bootstrap method for the simulation model of Section 2.6 in Figure 2.7. We compare the Monte Carlo estimate of MSE, defined in (2.7.1), (based on $N = 1,000$ samples of size $n = 1,000$) to the bootstrap MSE defined in (2.7.2) (based on a single sample of size $n = 1,000$ with $B = 10,000$). Figure 2.7 shows that the Monte Carlo MSE and the bootstrap MSE are in line, which illustrates the consistency of the method. The choice of the initial bandwidth does affect the size of the estimated MSE but not the behavior of the estimate and we conclude that this bootstrap algorithm can be used to select an optimal bandwidth parameter in the previously described method.

2.8 Illustration of the limit function F_α

In order to better understand the behavior of the limit function $F_{\boldsymbol{\alpha}}$ defined in (2.3.1) we calculate the analytical expression for the function $F_{\boldsymbol{\alpha}}$ in several models of the type (2.1.1). We consider the following one dimensional scenarios:

A. $X, T \sim U[0,2], f_0(e) = 384(e - 3/8)(5/8 - e)1_{[3/8,5/8]}(e)$ and $\alpha_0 = 0.5$.

B. $X, T \sim U[0,2], f_0(e) = \frac{\exp(-e)}{\exp(-3/8)-\exp(-3/5)}$(truncated exponential on [3/8,5/8], denoted by Exp[3/8,5/8]) and $\alpha_0 = 0.5$.

C. $X, T \sim$ Exp[0,20],$\varepsilon \sim U[5,10]$ and $\alpha_0 = 1$.

D. $X, T \sim$ Exp[0,20],$\varepsilon \sim$Exp[5,10] and $\alpha_0 = 1$.

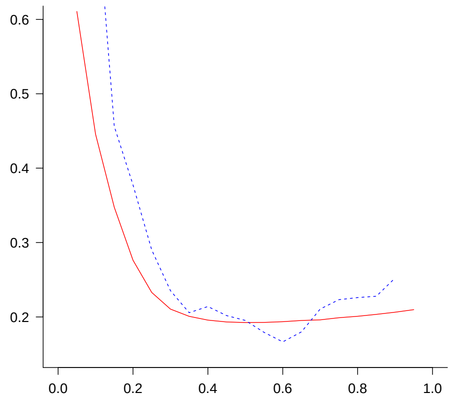


Figure 2.7: Estimated MSE(c) plot of $\hat{\alpha}_{3n}$ obtained from $N = 1,000$ Monte Carlo simulations (red, solid) and the bootstrap MSE for $c_0 = 0.25$ (blue, dashed) with $B = 10,000$, $n = 1,000$ and $\epsilon = 0.001$.

Scenario A corresponds to the simulation model of Section 2.6. We consider two different distributions (with compact support) for the observed variables X and T. The corresponding densities of the variables $T - \alpha X$ for different values of α is given in Figure 2.8a for uniform random variables X and T and in Figure 2.8b for truncated exponential random variables X and T. Figure 2.8 shows that the density $f_{T-\alpha X}$ is continuous and staying away from zero on the support of f_0 such that Assumption A3 given in Section 2.4.1 is satisfied in all four scenarios A-D. The above densities are moreover twice continuously differentiable except at a finite number of points and therefore Figure 2.8 also illustrates the plausibility of Assumption A5.

For each scenario we show figures of F_α for α in a neighborhood of α_0 in Figure 2.9. The function F_α is monotone and twice continuously differentiable with a strictly positive derivative on a truncated interval determined by the support of f_0 for each α considered in Figure 2.9. The function F_α converges to F_0 if the distance between α and α_0 decreases and we conclude that the assumptions in Section 2.4.1 are plausible, in especially for $\alpha \to \alpha_0$.

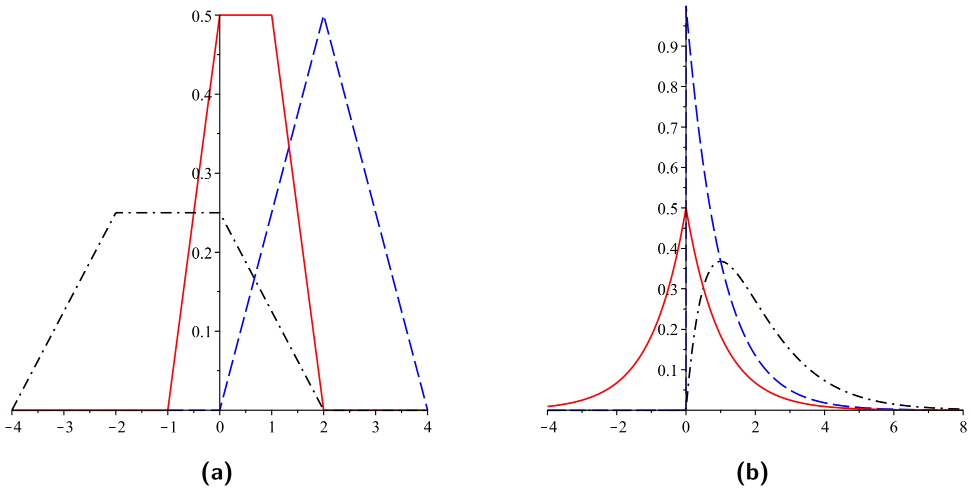


Figure 2.8: The density $f_{T-\alpha X}$ for (a) $X, T \sim U[0,2]$ and $\alpha = \alpha_0 = 0.5$ (red, solid), $\alpha = -1$ (blue, dashed) and $\alpha = 2$ (black, dashed-dotted) and (b) $X, T \sim$ Exp[0,20] and $\alpha = \alpha_0 = 1$ (red, solid), $\alpha = 0$ (blue, dashed) and $\alpha = -1$ (black, dashed-dotted)

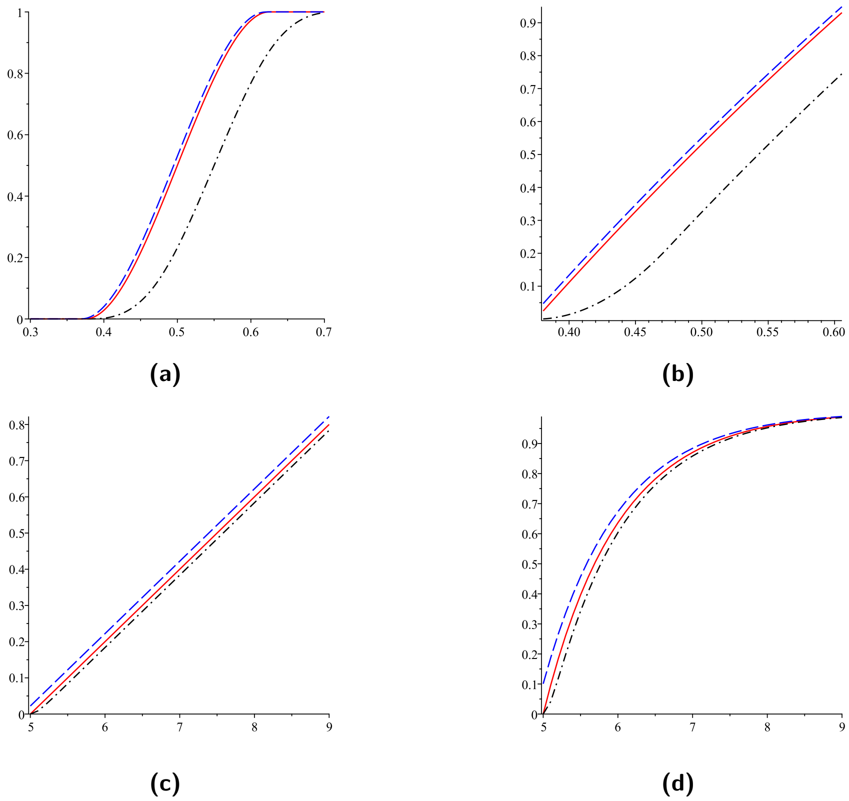


Figure 2.9: The function F_α for (a-b) scenario A and B with $\alpha = \alpha_0 = 0.5$ (red, solid), $\alpha = 0.505$ (blue, dashed) and $\alpha = 0.45$ (black, dashed-dotted) and (c-d) scenario C and D with $\alpha = \alpha_0 = 1$ (red, solid), $\alpha = 1.25$ (blue, dashed) and $\alpha = 0.85$ (black, dashed-dotted)

Chapter 3

Bootstrap procedures under current status data

Abstract

We study the behavior of two different bootstrap algorithms in the current status model and develop bootstrap procedures for constructing pointwise confidence intervals (CIs) for the distribution function F_0 of Y based on censored observations $(T_1, \Delta_1), \dots, (T_n, \Delta_n)$ from the random vector $(T, \Delta = 1_{\{Y \le T\}})$. In the first approach, we consider a smooth bootstrap procedure that is based on only resampling the censoring indicators Δ_i from a Bernoulli distribution with probability $\tilde{F}_{nh}(T_i)$, where $\tilde{F}_{nh}$ is the smoothed maximum likelihood estimator (SMLE) of the distribution function F_0. In the second approach we consider the nonparameteric bootstrap proposed by Efron (1979).

Asymptotic results show that, given the data, the L_2-distance between the bootstrap MLE and the underlying distribution function is of order $n^{-1/3}$. This result is in particular noteworthy for the nonparametric bootstrap given the fact that this bootstrap is inconsistent for generating the distribution of the MLE.

We construct pointwise CIs around the SMLE and prove the validity of interval estimation in the current status linear regression model for both bootstrap procedures. A comparison of both methods through simulation studies does not reveal a clear preference for one of both bootstrap approaches. The bandwidth parameter used in the smoothing procedure has a considerable influence on the behavior of the CIs. We show that a data-driven bandwidth parameter based on minimizing the mean squared error in combination with undersmoothing results in CIs with good coverage properties.

3.1 The MLE and the SMLE under current status data

Let $Z_1 = (T_1, \Delta_1), \ldots, Z_n = (T_n, \Delta_n)$ be the i.i.d. sample from the current status model described in Section 1.1 where the Y_i are interpreted as (nonnegative) survival times with distribution function F_0 and, instead of observing Y, a censoring variable $T \sim G$ is observed (with density g) independent of Y. We denote the probability measure of $Z = (T, \Delta)$ by P_0. The density of Z is given by

$$p_{F_0}(t, \delta) = [\delta F_0(t) + (1-\delta)\{1 - F_0(t)\}]\, g(t).$$

An important property of the MLE, defined in (1.1.2) as the maximizer of the log likelihood over all possible distribution functions, is the so-called *switch relation*: Let $\mathbb{G}_n$ be the empirical distribution function of $T_1, \ldots, T_n$ and define the process V_n by

$$V_n(t) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \Delta_i 1_{\{T_i \le t\}}, \tag{3.1.1}$$

and the process U_n by

$$U_n(a) \stackrel{\text{def}}{=} \text{argmin}\{t \in \mathbb{R} : V_n(t) - a\mathbb{G}_n(t)\}. \tag{3.1.2}$$

Then we get the *switch relation*

$$\hat{F}_n(t) \ge a \iff U_n(a) \le t, \tag{3.1.3}$$

see also Figure 3.1. As a consequence of the switch relation, the study of the MLE $\hat{F}_n$ can

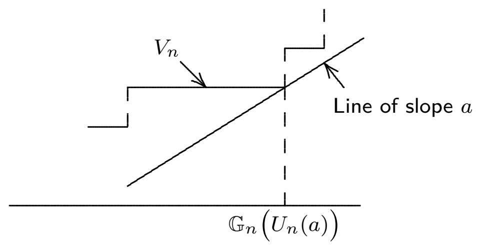


Figure 3.1: The switch relation.

be reduced to the study of the inverse process U_n, taking $a = F_0(t)$. From Groeneboom and Jongbloed (2014) we have the following result:

Theorem 3.1.1 (Groeneboom and Jongbloed, 2014, Theorem 11.3)**.** *Suppose F_0 has a continuous density f_0 with support [0,M] that satisfies,*

$$0 < \inf_{t \in [0,M]} f_0(t) < \sup_{t \in [0,M]} f_0(t) < \infty.$$

Also suppose that the observation distribution G has a continuous derivative g that stays away form zero and infinity on $[0, M]$. Let U_n be defined by (3.1.2) and let

$$U_0(a) \stackrel{\text{def}}{=} F_0^{-1}(a),$$

for every $a \in [0,1]$. Then there are positive constants K_1 and K_2, such that, for all $a \in (0,1)$ and for all $x > 0$,

$$P_0\left\{n^{1/3}\left|U_n(a) - U_0(a)\right| \geq x\right\} \leq K_1 e^{-K_2 x^3}. \tag{3.1.4}$$

It follows from Theorem 3.1.1 and the *switch relation* (3.1.3) that there exists a positive constant $K > 0$ such that

$$\mathbb{E}\left|\hat{F}_n(t) - F_0(t)\right|^p \leq K n^{-p/3} \qquad \text{for all } t \in (0, M). \tag{3.1.5}$$

This can be seen by nothing that

$$\mathbb{E}\left[n^{1/3}\{\hat{F}_n(t) - F_0(t)\}_+\right]^p = \int_0^\infty P_0\left\{n^{1/3}\{\hat{F}_n(t) - F_0(t)\} \geq x\right\} p x^{p-1} dx,$$

where $\{\hat{F}_n(t) - F_0(t)\}_+$ denotes the positive part of $\{\hat{F}_n(t) - F_0(t)\}$ and that

$$P_0\left\{U_n\left(a + n^{-1/3}x\right) \leq t\right\} = P_0\left[n^{1/3}\left\{U_n\left(a + n^{-1/3}x\right) - U_0\left(a + n^{-1/3}x\right)\right\} \leq n^{1/3}\left\{t - U_0\left(a + n^{-1/3}x\right)\right\}\right].$$

In this chapter we show that a result analogous to the result of Theorem 3.1.1 holds in a bootstrap sample obtained from the original sample $Z_1, \ldots, Z_n$. In Section 3.2 and Section 3.3 we consider two bootstrap samples, generated by the smooth respectively nonparametric bootstrap (See Section 1.1.2) and we show how this result is used in proving the validity of the bootstrap for constructing pointwise CIs for the distribution function $F_0(t)$ when t is an interior point of $[0, M]$. We estimate $F_0(t)$ by the Smoothed MLE obtained by first estimating the MLE $\hat{F}_n$ and then smoothing this using a smoothing kernel, i.e.

$$\tilde{F}_{nh}(t) \stackrel{\text{def}}{=} \int \mathbb{K}\left(\frac{t-x}{h}\right) d\hat{F}_n(x), \tag{3.1.6}$$

for some bandwidth $h > 0$, where $\mathbb{K}$ is the integrated kernel (1.1.6) (see Section 1.1.2). As argued in Section 11.2 in Groeneboom and Jongbloed (2014), the asymptotic normality of the SMLE, stated in Theorem 1.1.2, can be proved using that

$$\int_{t-h}^{t+h}\left\{\hat{F}_n(x) - F_0(x)\right\}^2 dx = O_p(hn^{-2/3}),$$

which follows from (3.1.5) and an application of Markov's inequality. Note that this last statement does not follow from a global L_2-distance of order $n^{-1/3}$ between the MLE $\hat{F}_n$ and the true distribution function F_0 and therefore, it is a refinement of the usual Hellinger distance calculations.

Similarly to the notations used in the previous Chapter 2, we use the notations K_h and $\mathbb{K}_h$ to denote the scaled versions of K and $\mathbb{K}$ respectively, given by

$$K_h(\cdot) = h^{-1}K(\cdot/h) \quad \text{and} \quad \mathbb{K}_h(\cdot) = \mathbb{K}(\cdot/h).$$

3.2 Pointwise confidence intervals using the smooth bootstrap

We obtain a bootstrap sample $(T_1, \Delta_1^*), \ldots, (T_n, \Delta_n^*)$ by keeping the T_i in the original sample fixed and by resampling the Δ_i^* from a Bernoulli distribution with probability $\tilde{F}_{nh}(T_i)$. This procedure is referred to as the smooth bootstrap algorithm. The following bootstrap $1-\alpha$ interval is suggested:

$$\left[\tilde{F}_{nh}(t) - Z_{1-\alpha/2}^*(t), \tilde{F}_{nh}(t) - Z_{\alpha/2}^*(t)\right], \tag{3.2.1}$$

where $Z_\alpha^*(t)$ is the αth quantile of B values of

$$Z_{nh}(t) \stackrel{\text{def}}{=} \tilde{F}_{nh}^*(t) - \int \mathbb{K}_h(t-u)\, d\tilde{F}_{nh}(u),$$

where B is the number of bootstrap samples. Here $\tilde{F}_{nh}^*(t)$ is the SMLE in the bootstrap sample $(T_1, \Delta_1^*), \ldots, (T_n, \Delta_n^*)$ defined in the same way as in (3.1.6) but with $\hat{F}_n$ replaced by $\hat{F}_n^*$, i.e. the MLE in the bootstrap sample.
We have the following main result:

$$n^{2/5}\left\{\tilde{F}_{nh}^*(t) - \int \mathbb{K}_h(t-u)\, d\tilde{F}_{nh}(u)\right\} \xrightarrow{d} N(0, \sigma^2(t)), \tag{3.2.2}$$

given the data $Z_1, \ldots, Z_n$, almost surely along sequences $Z_1, Z_2, \ldots$, where σ is defined in Theorem 1.1.2 as

$$\sigma^2(t) \stackrel{\text{def}}{=} \frac{F_0(t)\{1-F_0(t)\}}{cg(t)} \int K(u)^2\, du. \tag{3.2.3}$$

The proof of (3.2.2) is given in Appendix B, Section B.1.2 and requires first proving a bootstrap analogue to result (3.1.4) which we describe in more detail below. We first introduce the following notations: Let $\tilde{\mathbb{P}}_n$ denote the empirical measure of $(T_1, \Delta_1^*), \ldots, (T_n, \Delta_n^*)$. We write

$$\frac{1}{n}\sum_{i=1}^n f(T_i, \Delta_i^*) = \int f(u, \delta^*)\, d\tilde{\mathbb{P}}_n(u, \delta^*),$$

for some bounded function $f : [0, M] \times \{0, 1\} \to \mathbb{R}$. Note that for any bounded function $h : [0, M] \to \mathbb{R}$ holds that

$$\frac{1}{n}\sum_{i=1}^{n} h(T_i) = \int h(u)\, d\tilde{\mathbb{P}}_n(u, \delta^*) = \int h(u)\, d\mathbb{G}_n(u).$$

Furthermore, let $\tilde{P}_n$ (respectively $\tilde{E}_n$) denote the conditional probability measure (respectively conditional expectation), given $Z_1, \ldots, Z_n$ and note that

$$\tilde{P}_n\left(\Delta_i^* = 1\right) = \tilde{F}_{nh}(T_i) \quad 1 \le i \le n.$$

We next define the process

$$\tilde{V}_n(t) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \Delta_i^* 1_{\{T_i \le t\}} = \int_{u\in[0,t]} \delta^*\, d\tilde{\mathbb{P}}_n(u, \delta^*),$$

and the processes

$$\tilde{U}_0(a) \stackrel{\text{def}}{=} \text{argmin}\{t : \tilde{F}_{nh}(t) \ge a\},$$

and

$$\tilde{U}_n(a) \stackrel{\text{def}}{=} \text{argmin}\{t \in \mathbb{R} : \tilde{V}_n(t) - a\mathbb{G}_n(t)\}.$$

We have the following result which is the smooth bootstrap version of Theorem 3.1.1.

Lemma 3.2.1. *There are positive constants $\tilde{K}_1$ and $\tilde{K}_2$, such that, almost surely, for all $x > 0$ and all large n we have that*

$$\tilde{P}_n\left\{n^{1/3}\left|\tilde{U}_n(a) - \tilde{U}_0(a)\right| \ge x\right\} \le \tilde{K}_1 e^{-\tilde{K}_2 x^3}.$$

Using the *smooth bootstrapped switch relation* with $a_n = \tilde{F}_{nh}(t)$,

$$\tilde{P}_n\left\{n^{1/3}\{\hat{F}_n^*(t) - \tilde{F}_{nh}(t)\} \ge x\right\} = \tilde{P}_n\left\{\tilde{U}_n(a_n + n^{-1/3}x) \le t\right\},$$

we get by Lemma 3.2.1, that there exists a positive constant $\tilde{K}$ such that the indicator of the set

$$\left\{\exists t \in [0, M] : \tilde{E}_n\left|\hat{F}_n^*(t) - \tilde{F}_{nh}(t)\right|^p > \tilde{K}n^{-p/3}\right\},$$

is zero almost surely.

Note that we subtract the convolution SMLE $\int \mathbb{K}_h(t-u)\, d\tilde{F}_{nh}(u)$ using the original data instead of the SMLE $\tilde{F}_{nh}(t)$ itself in the definition of $Z_{nh}(t)$ due to the bias of the SMLE. This is in line with the method proposed by Sen and Xu (2015) where the authors subtract the SMLE instead of the MLE of the original data for constructing CIs around the MLE (see Section 1.1.2). One needs to introduce an additional level of smoothing in order to construct valid intervals using the smooth bootstrap procedure. Note that we can write:

$$\int \mathbb{K}_h(t-u)\, d\tilde{F}_{nh}(u) = \int \mathbb{K}_h(t-u) \left\{ \int K_h(u-v) d\hat{F}_n(v) \right\} du.$$
$$= \int \int \mathbb{K}\{(t-v)/h - w\} K(w)\, dw d\hat{F}_n(v).$$

In practice we therefore have to compute the convolution kernel $\widetilde{\mathbb{K}}$, defined by

$$\widetilde{\mathbb{K}}(x) \stackrel{\text{def}}{=} \int \mathbb{K}(x-w) K(w)\, dw. \tag{3.2.4}$$

A picture of the functions K, $\mathbb{K}$ and $\widetilde{\mathbb{K}}$ is given in Figure 3.2 using the triweight kernel given by

$$K(u) = \frac{35}{32} \left(1-u^2\right)^3 1_{[-1,1]}(u),$$

and the Epanechnikov kernel given by

$$K(u) = \tfrac{3}{4}(1-u^2) 1_{[-1,1]}(u).$$

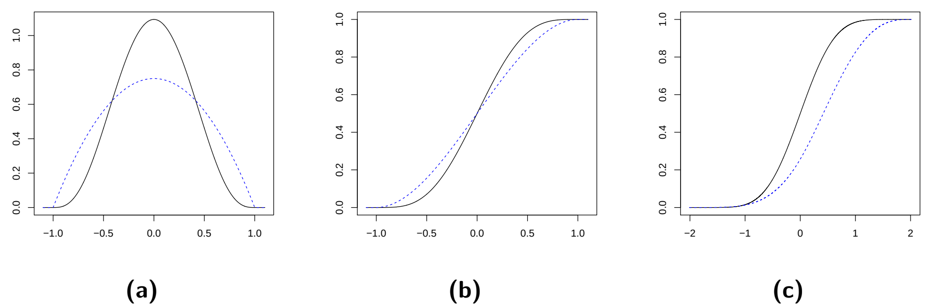


Figure 3.2: (a) kernel K, (b) integrated kernel $\mathbb{K} : x \mapsto \int_{-\infty}^{x} K(w)\, dw$ and (c) convolution kernel $\widetilde{\mathbb{K}} : x \mapsto \int \mathbb{K}(x-w)K(w)\, dw$ for the triweight kernel $K : x \mapsto \frac{35}{32}(1-x^2)^3 1_{[-1,1]}(x)$ (black, solid) and Epanechnikov kernel $K : x \mapsto \frac{3}{4}(1-x^2)1_{[-1,1]}(x)$ (blue, dashed).

In the remainder of this section, we describe techniques to improve the CIs introduced in (3.2.1) by (a) considering Studentized CIs through estimation of the variance σ^2 in (3.2.3) and (b) taking into account the boundary effects of the kernel estimates. In practice, one should also correct for the bias defined in Theorem 1.1.2 when constructing CIs around the SMLE by considering intervals of the type

$$\left[\tilde{F}_{nh}(t) - Z^*_{1-\alpha/2}(t) - \beta(t) n^{-2/5}, \tilde{F}_{nh}(t) - Z^*_{\alpha/2}(t) - \beta(t) n^{-2/5} \right],$$

where

$$\beta(t) \stackrel{\text{def}}{=} \frac{c^2 f_0'(t)}{2} \int u^2 K(u)\, du. \tag{3.2.5}$$

The boundary issue will be discussed in more detail in Section 3.4 where we will illustrate how a data-driven bandwidth selection procedure can be used to reduce the bias effect present in the CIs given in (3.2.1).

3.2.1 Studentized confidence intervals

Usually the performance of the bootstrap CIs works best if one uses a pivot, obtained by Studentizing. In each bootstrap sample we therefore estimate the variance σ^2 defined in (3.2.3), apart from the factor $cg(t)$, which drops out in the Studentized bootstrap procedure, by

$$S^*_{nh}(t) \stackrel{\text{def}}{=} n^{-2} \sum_{i=1}^{n} K_h(t-T_i)^2 \left(\Delta^*_i - \hat{F}^*_n(T_i)\right)^2. \tag{3.2.6}$$

The variance estimate defined in (3.2.6) is inspired by the fact that the SMLE $\tilde{F}_{nh}$ is asymptotically equivalent to the toy estimator

$$\tilde{F}^{toy}_{nh}(t) \stackrel{\text{def}}{=} \int \mathbb{K}_h(t-x)\, dF_0(x) + \frac{1}{n}\sum_{i=1}^{n} \frac{K_h(t-T_i)\{\Delta_i - F_0(T_i)\}^2}{g(T_i)}, \tag{3.2.7}$$

which has sample variance

$$S_n(t) \stackrel{\text{def}}{=} \frac{1}{n^2}\sum_{i=1}^{n} \frac{K_h(t-T_i)^2\left(\Delta_i - F_0(T_i)\right)^2}{g(T_i)^2}. \tag{3.2.8}$$

We next compute

$$Q^*_{nh}(t) \stackrel{\text{def}}{=} \frac{\tilde{F}^*_{nh}(t) - \int \mathbb{K}_h(t-u)\, d\tilde{F}_{nh}(u)}{\sqrt{S^*_{nh}(t)}}.$$

Let $Q^*_\alpha(t)$ be the αth quantile of B values of $Q^*_{nh}(t)$. Then the following bootstrap $1-\alpha$ interval is suggested:

$$\left[\tilde{F}_{nh}(t) - Q^*_{1-\alpha/2}(t)\sqrt{S_{nh}(t)}, \tilde{F}_{nh}(t) - Q^*_{\alpha/2}(t)\sqrt{S_{nh}(t)}\right], \tag{3.2.9}$$

where $S_{nh}(t)$ is the variance estimate in the original sample obtained by replacing $\Delta^*_i - \hat{F}^*_n(T_i)$ in (3.2.6) by $\Delta_i - \hat{F}_n(T_i)$. Note that we do not need an estimate of the density g in each of the observations T_i as a consequence of the fact that $g(u)$ is close to $g(t)$ for $u \in [t-h, t+h]$. On the contrary, estimation of g is inevitable if one wants to consider Wald-type CIs for the distribution function based on the asymptotic normality result of the SMLE. More details of the construction of Wald type CIs in given in the simulations presented in Section 3.4.

3.2.2 Boundary correction

It is well-known that kernel density and distribution estimators without boundary correction are generally inconsistent at the boundary of the support $[0, M]$. We therefore use the boundary correction method proposed in Groeneboom and Jongbloed (2014), and define the SMLE as

$$\tilde{F}_{nh}^{(bc)}(t) \stackrel{\text{def}}{=} \int \left\{ \mathbb{K}\left(\frac{t-x}{h}\right) + \mathbb{K}\left(\frac{t+x}{h}\right) - \mathbb{K}\left(\frac{2M-t-x}{h}\right) \right\} d\hat{F}_n(x). \quad (3.2.10)$$

The boundary corrected version of $Z_{nh}^*(t)$ is defined by

$$Z_{nh}^{(bc)*}(t) \stackrel{\text{def}}{=} \tilde{F}_{nh}^{(bc)*}(t) - \int \{\mathbb{K}_h(t-x) + \mathbb{K}_h(t+x) - \mathbb{K}_h(2M-t-x)\} \, d\tilde{F}_{nh}^{(bc)}(x).$$

It is straightforward to show that the asymptotic normality result (3.2.2) remains valid under this boundary correction. We also have the following lemma.

Lemma 3.2.2. *Let the boundary corrected estimate $\tilde{F}_{nh}^{(bc)}$ be defined by (3.2.10), and let $\widetilde{\mathbb{K}}_h$ be defined by*

$$\widetilde{\mathbb{K}}_h(u) = \widetilde{\mathbb{K}}(u/h), \qquad u \in \mathbb{R},$$

where the convolution kernel $\widetilde{\mathbb{K}}$ is defined by (3.2.4). Moreover, let $0 < h \le M/3$. Then:

$$\begin{aligned} &\int \{\mathbb{K}_h(t-x) + \mathbb{K}_h(t+x) - \mathbb{K}_h(2M-t-x)\} \, d\tilde{F}_{nh}^{(bc)}(x) \\ &= \int \left\{\widetilde{\mathbb{K}}_h(t-x) + \widetilde{\mathbb{K}}_h(t+x) - \widetilde{\mathbb{K}}_h(2M-t-x)\right\} d\hat{F}_n(x). \end{aligned} \quad (3.2.11)$$

From Lemma 3.2.2, it follows that we can write

$$\begin{aligned} Z_{nh}^{(bc)*}(t) &= \int \{\mathbb{K}_h(t-x) + \mathbb{K}_h(t+x) - \mathbb{K}_h(2M-t-x)\} \, d\big(\hat{F}_n^* - \tilde{F}_{nh}^{(bc)}\big)(x) \\ &= \int \{\mathbb{K}_h(t-x) + \mathbb{K}_h(t+x) - \mathbb{K}_h(2M-t-x)\} \, d\hat{F}_n^*(x) \\ &\quad - \int \left\{\widetilde{\mathbb{K}}_h(t-x) + \widetilde{\mathbb{K}}_h(t+x) - \widetilde{\mathbb{K}}_h(2M-t-x)\right\} d\hat{F}_n(x). \end{aligned}$$

The proof of Lemma 3.2.2 is given in Appendix B, Section B.1. A picture of the MLE, together with the SMLE, both corrected and uncorrected for boundary effects is shown in Figure 3.3a for a sample from the truncated exponential distribution on $[0, 2]$ (See Section 3.4 for a detailed description of the model). Figure 3.3b presents the boundary corrected and uncorrected convolution SMLE and clearly shows the improvement of the boundary correction. From this point onwards, we will work with the boundary corrected SMLE and use the notation $\tilde{F}_{nh}(t)$ whenever we refer to the SMLE.

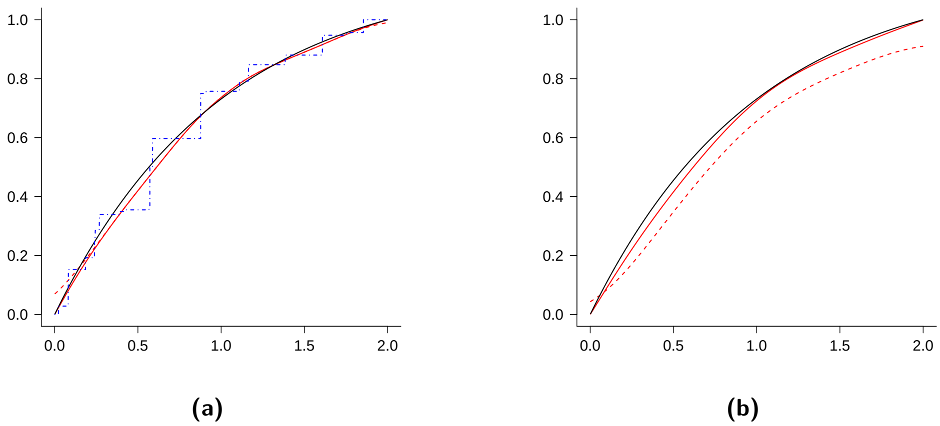


Figure 3.3: Truncated exponential samples: (a) F_0 (black, solid), MLE $\hat{F}_n$ (blue, dashed-dotted), SMLE $\tilde{F}_{nh}^{(bc)}$ with boundary correction (red, solid) and SMLE $\tilde{F}_{nh}$ without boundary correction (red, dashed) and (b) F_0 (black, solid), convolution SMLE with boundary correction (red, solid) and convolution SMLE without boundary correction (red, dashed); $n = 1,000$ and $h = 2n^{-1/5}$.

3.3 Pointwise confidence intervals using the nonparametric bootstrap

In this section we consider the nonparametric bootstrap proposed by Efron (1979) and generate a bootstrap sample $(T_1^*, \Delta_1^*), \ldots, (T_n^*, \Delta_n^*)$ by resampling with replacement from the original sample. Denote the empirical probability measure of $Z_1, \ldots, Z_n$ by $\mathbb{P}_n$. The bootstrap empirical measure is

$$\hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} M_{ni} 1_{Z_i},$$

where 1_{Z_i} denotes the point mass at $Z_i = (T_i, \Delta_i)$ and

$$M_n = (M_{n1}, \ldots, M_{nn}) \sim \text{multinomial}(n, n^{-1}, \ldots, n^{-1}),$$

is a vector of multinomial weights, independent of $Z_1, \ldots, Z_n$. Let $\hat{P}_n$ denote the conditional probability measure w.r.t. the weights given the sample $Z_1, \ldots, Z_n$ and define the process

$$\hat{U}_n(a) \stackrel{\text{def}}{=} \text{argmin}\{t \in [0, M] : \hat{V}_n(t) - a\hat{\mathbb{G}}_n(t)\} \qquad 0 < a < 1,$$

with processes $\hat{V}_n$ and $\hat{\mathbb{G}}_n$ defined by

$$\hat{V}_n(t) \stackrel{\text{def}}{=} \int_{u\in[0,t]} \delta\, d\hat{\mathbb{P}}_n(u,\delta) \quad \text{and} \quad \hat{\mathbb{G}}_n(t) \stackrel{\text{def}}{=} \int_{u\in[0,t]} d\hat{\mathbb{P}}_n(u,\delta) \quad t \in [0, R]. \tag{3.3.1}$$

Lemma 3.3.1. *There are positive constants $\hat{K}_1$ and $\hat{K}_2$, such that for all large n we have that*

$$\left\{\exists x \in [0, M] : \hat{P}_n \left\{n^{1/3} \left|\hat{U}_n(a) - U_0(a)\right| \geq x\right\} > \hat{K}_1 e^{-\hat{K}_2 x^{3/2}}\right\} = o_p(1).$$

Lemma 3.3.1 implies that the probability that for all $x \in [0, M]$, and $a = F_0(t)$,

$$\hat{P}_n \left\{n^{1/3} \left|\hat{U}_n(a) - U_0(a)\right| \geq x\right\} \leq K_1 e^{-K_2 x^{3/2}},$$

tends to 1 as $n \to \infty$. The proof of Lemma 3.3.1 is given in Appendix B, Section B.2.1. The proof uses empirical process theory and results on tail probabilities for $\|\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)\|_{\mathcal{F}}$ for classes $\mathcal{F}$ with finite entropy integrals (see van der Vaart and Wellner (1996), p.81 for the notation $\| \cdot \|\|_{\mathcal{F}}$). The analogue results in Theorem 3.1.1 and Lemma 3.2.1 are proved using martingale theory in Section 11.2 of Groeneboom and Jongbloed (2014) for the original sample and in Appendix B, Section B.1.1 for a smooth bootstrap empirical process respectively. It follows from Lemma 3.3.1 and the *bootstrapped switch relation* given by

$$\hat{P}_n \left\{n^{1/3}\{\hat{F}_n^*(t) - F_0(t)\} \geq x\right\} = \hat{P}_n \left\{\hat{U}_n \left(a + n^{-1/3}x\right) \leq t\right\},$$

that there exists a positive constant $\hat{K} > 0$ such that,

$$\left\{\exists t \in [0, M] : \hat{E}_n \left|\hat{F}_n^*(t) - F_0(t)\right|^p > \hat{K} n^{-p/3}\right\} = o_p(1),$$

where $\hat{E}_n$ denotes the conditional expectation given $Z_1, \ldots, Z_n$.

We now continue with the construction of the pointwise CIs. Let $\tilde{F}_{nh}^*(t)$ be the bootstrapped SMLE based on replacing $\hat{F}_n$ in (3.1.6) by the bootstrapped MLE $\hat{F}_n^*$. Then we have, using Lemma 3.3.1, the following result

$$n^{2/5}\left\{\tilde{F}_{nh}^*(t) - \tilde{F}_{nh}(t)\right\} \xrightarrow{d} N(0, \sigma^2), \tag{3.3.2}$$

given the data $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$, in probability, where σ^2 is defined by (3.2.3). The proof of (3.3.2) is given in Appendix B, Section B.2.2. A bootstrap $100(1-\alpha)\%$ interval is next given by (3.2.9) with $Q_\alpha(t)$ replaced by the αth quantile of B values of

$$\frac{\tilde{F}_{nh}^*(t) - \tilde{F}_{nh}(t)}{\sqrt{S_{nh}^*(t)}},$$

where $S_{nh}^*(t)$ now represents the variance estimate in the nonparametric bootstrap sample. Note that, in contrast to the smooth bootstrap method described in Section 3.2, we do not longer need to estimate the convolution SMLE.

3.4 Simulations

Two different simulation models are considered in this section to study the finite sample behavior of the CIs introduced in Section 3.2 and Section 3.3. The effect of the choice of the kernel function on the performance of the intervals is addressed in the first simulation study. A comparison between the proposed SMLE-based intervals and the MLE-based intervals of Banerjee and Wellner (2005) and Sen and Xu (2015) and a discussion of the construction of Wald-type CIs using the quantiles of the asymptotic normal distribution of the SMLE is also considered. The bias issues that arise when constructing CIs around the (biased) SMLE instead of the (unbiased) MLE are illustrated in the second simulation study. A data-driven bandwidth selection procedure is proposed and undersmoothing the bandwidth is used to take the bias problems into account.

3.4.1 Simulation study 1: Comparing CIs for the distribution function under current status data

In the first simulation setting both the event times and censoring times are sampled from a Uniform(0,2)-distribution. Since the derivative of the uniform density equals zero, the bias defined in (3.2.5) equals zero thence the SMLE is an unbiased estimate of the uniform distribution function. No bias correction is needed in the construction of the CIs.

For sample sizes $n = 100; 500; 1,000$ and $2,000$ we generated $5,000$ data sets from this uniform model. The number of bootstrap samples within each simulation run equals $B = 1,000$. We use the bandwidth $h = cn^{-1/5}$, where the constant $c = 2$ corresponds to the length of the interval $[0, 2]$. Two different choices for the kernel, the triweight kernel and the Epanechnikov kernel, are considered. The results for our SMLE-based CIs (3.2.9) and for the MLE-based methods of Banerjee and Wellner (2005) and Sen and Xu (2015) are compared. Table 3.1 shows the coverage percentage, i.e. the number of times (out of the 5,000 simulation runs) that $F_0(t)$ is in the 95% CIs, and the average length of the 95% CIs around $F_0(t)$ for the uniform model and $t = 1$. For the Studentized SMLE-based CIs, the results in Table 3.1 correspond to the results for the smooth bootstrap.

For each point $t = 0.02, 0.04, \ldots, 2$, Figure 3.4a presents the proportion of times that $F_0(t)$ is not in the 95% CIs, for the Studentized SMLE-based CIs obtained with the smooth bootstrap procedure (3.2.9) using the Epanechnikov kernel and the triweight kernel and illustrates that the choice of the kernel has only a small effect on the coverage proportions. The comparison between the CIs defined in (3.2.1) and the Studentized CIs defined in (3.2.9) shown in Figure 3.4b reveals that the non-Studentized SMLE-based CIs are slightly anti-conservative near the left boundary of the interval and have a coverage that is less good than the Studentized CIs. In the rest of the simulation section, we therefore use the

Table 3.1: Uniform samples: Average length (L) and coverage proportion (CP) of the smooth SMLE-based CIs and the MLE-based CIs proposed by Banerjee and Wellner (2005) and Sen and Xu (2015) around $F_0(1)$ for different choices of the kernel and different sample sizes n. ($h = 2n^{-1/5}$, $\alpha = 0.05$)

	Studentized SMLE-based CI				Banerjee-Wellner		Sen-Xu			
	Triweight		Epanechnikov				Triweight		Epanechnikov	
n	CP	L	CP	L	CP	L	CP	L	CP	L
100	0.9674	0.2799	0.9642	0.2376	0.9514	0.3897	0.9432	0.4620	0.9530	0.4625
500	0.9528	0.1473	0.9546	0.1276	0.9496	0.2311	0.9364	0.2532	0.9420	0.2536
1,000	0.9374	0.1072	0.9400	0.0928	0.9502	0.1846	0.9346	0.2024	0.9404	0.2028
2,000	0.9506	0.0827	0.9498	0.0710	0.9586	0.1466	0.9484	0.1598	0.9518	0.1599

Studentized CIs, referred to as the SMLE-based CIs.

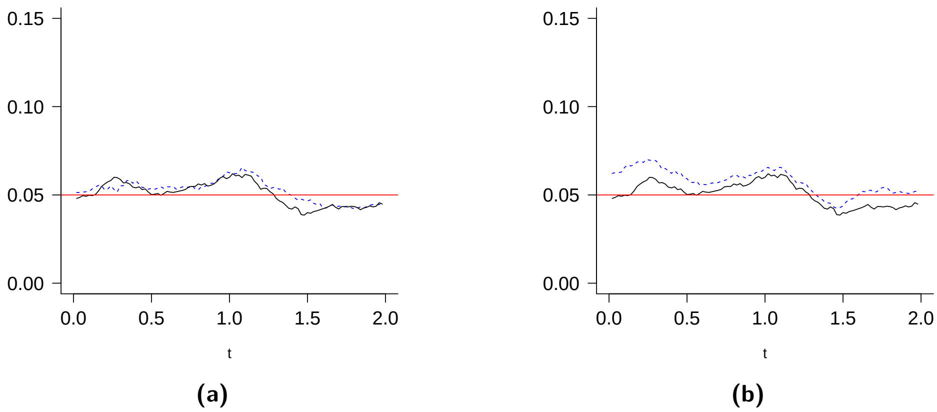


Figure 3.4: Uniform samples: Proportion of times that $F_0(t)$, $t = 0.02, 0.04, \ldots$ is not in the 95% CIs in $N = 5,000$ samples using $B = 1,000$ smooth bootstrap samples. (a) Studentized SMLE-based CIs (3.2.9) with the Epanechnikov kernel (black, solid) and the triweight kernel (blue, dashed) and (b) Studentized SMLE-based CIs (3.2.9) (black, solid) and classical SMLE-based CIs (3.2.1) (blue, dashed) with the Epanechnikov kernel. $n = 1,000$ and $h = 2n^{-1/5}$.

The performances of the SMLE-based CIs, illustrated in Figure 3.5a, are comparable. The bootstrap intervals based on the nonparametric bootstrap procedure avoid however calculation of the convolution SMLE. The MLE-based CIs obtained via the LR-method

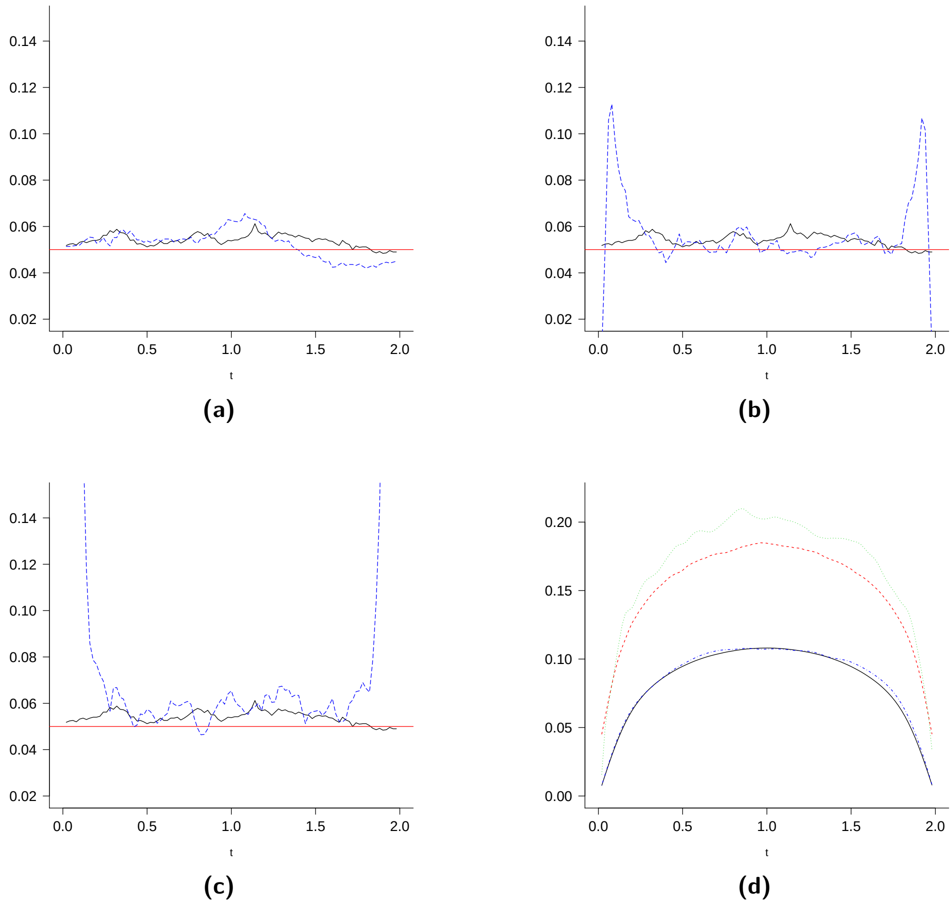
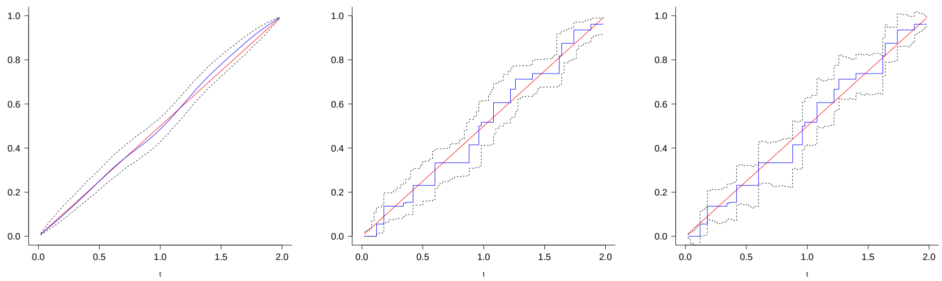


Figure 3.5: Uniform samples: Proportion of times that $F_0(t)$, $t = 0.02, 0.04, \ldots$ is not in the 95% CIs in $N = 5,000$ samples using the triweight kernel and $B = 1,000$ bootstrap samples for (a) SMLE-based CIs using the nonparametric (black, solid) and smooth (blue, dashed) bootstrap, (b) Banerjee and Wellner (2005) CIs (blue, dashed) and SMLE-based nonparametric bootstrap CIs (black, solid) and (c) Sen and Xu (2015) CIs (blue, dashed) and SMLE-based nonparametric bootstrap CIs (black, solid). (d) The average lengths of the SMLE-based nonparametric (black, solid) and smooth (blue, dashed-dotted) bootstrap CIs and the MLE-based CIs of Banerjee and Wellner (2005) (red, dashed) and of Sen and Xu (2015) (green, dotted). $n = 1,000$ and $h = 2n^{-1/5}$.

of Banerjee and Wellner (2005) (Figure 3.5b) and the smooth bootstrap method of Sen and Xu (2015) (Figure 3.5c) have similar coverage proportions in the middle of the interval $[0, 2]$ but have a worse behavior near the boundaries of the interval compared to the SMLE-based intervals. Figure 3.5d shows the average length of both bootstrap intervals around the SMLE in comparison with the average length of the LR CIs of Banerjee and Wellner (2005) and the smooth MLE-based CIs of Sen and Xu (2015). The length of the MLE-based intervals is larger than the length of the SMLE-based intervals due to the fact that the MLE converges at the slower rate $n^{1/3}$.

The CIs for one sample of size $n = 1,000$ are shown in Figure 3.6. The MLE-based CIs of Sen and Xu (2015) do not have monotone bounds. One may wonder if one really wants to use the MLE for estimating the distribution function if one resamples from the SMLE as in Sen and Xu (2015) since one uses smoothness conditions that allow for estimating the distribution function at a faster rate than the convergence rate of the MLE. The pointwise CIs around the SMLE change smoothly over the interval whereas MLE-based intervals change in discrete steps.



(a) **(b)** **(c)**

Figure 3.6: Uniform samples: F_0 (red solid). (a) Studentized SMLE-based CI, (b) Banerjee and Wellner (2005) CI and (c) Sen and Xu (2015) CI based on one sample of size $n = 1,000$ using $B = 1,000$ smooth bootstrap samples. In (a) the SMLE (blue, solid) using the triweight kernel is given and in (b,c) the MLE (blue, step function) is given. $h = 2n^{-1/5}$.

Instead of constructing the Studentized bootstrap intervals where the quantiles of the limiting distribution of the SMLE are derived from the bootstrap distribution, one can consider Wald-type CIs using the quantiles of the normal distribution and an estimate of the asymptotic variance. We compare three different estimates $\hat{\sigma}_{nh}$ for σ defined in (3.2.3) and construct CIs given by:

$$\Big[\tilde{F}_{nh}(t) - z_{1-\alpha/2}(n^{-2/5}\hat{\sigma}_{nh}(t)) - n^{-2/5}\beta(t);\quad \tilde{F}_{nh}(t) - z_{\alpha/2}(n^{-2/5}\hat{\sigma}_{nh}(t)) - n^{-2/5}\beta(t)\Big], \tag{3.4.1}$$

where z_α is the αth quantile of the standard normal distribution. The bias term β, defined in (3.2.5), is zero in this simulation study. The effect of β on the behavior of the intervals will be discussed in the second simulation study below.

A first estimate for $\hat{\sigma}_{nh}$ is given by

$$\hat{\sigma}^2_{1,nh}(t) \stackrel{\text{def}}{=} \frac{F_n(t)\{1-F_n(t)\}}{cg_{nh}(t)} \int K(u)^2\,du, \tag{3.4.2}$$

where g_{nh} is a classical kernel estimate for the density g of the observation time $T \sim U(0,2)$, using again the triweight kernel with bandwidth $h = 2n^{-1/5}$.

A second estimate for σ is inspired by the fact that the SMLE is asymptotically equivalent to the toy-estimator defined in (3.2.7), which has a sample variance given in (3.2.8). This suggests taking the second estimate $n^{-2/5}\hat{\sigma}_{2,nh}$ equal to the root of (3.2.8) where F_0 is replaced by the MLE $\hat{F}_n$ and g is replaced by the kernel density estimate g_{nh}, i.e.

$$n^{-4/5}\hat{\sigma}^2_{2,nh}(t) \stackrel{\text{def}}{=} \frac{1}{n^2}\sum_{i=1}^{n} \frac{K_h(t-T_i)^2\left(\Delta_i - \hat{F}_n(T_i)\right)^2}{g_{nh}(T_i)^2}. \tag{3.4.3}$$

Contrary to the bootstrap procedure for constructing CIs defined in (3.2.9), both estimates $\hat{\sigma}_{1,nh}$ and $\hat{\sigma}_{2,nh}$ require estimating the density g. A bootstrap based estimate for the variance is finally given by

$$\hat{\sigma}^2_{3,nh}(t) \stackrel{\text{def}}{=} \frac{1}{B}\sum_{b=1}^{B} \left(\tilde{F}^b_{nh}(t) - \tilde{F}_{nh}(t)\right)^2, \tag{3.4.4}$$

where $\tilde{F}^b_{nh}(t)$ is the SMLE in the bth bootstrap run. We use the nonparametric bootstrap to estimate $\sigma^2_{3,nh}$ in our simulations. Figure 3.7 compares the proportion of times that $F_0(t)$ is not in the 95% CIs, between the nonparametric bootstrap CIs with the Wald-type CIs in (3.4.1) using the triweight kernel and the three different variance estimates described above. Pointwise confidence bands for the variance estimates are illustrated in Figure 3.8. The curves show the average variance estimate and the 5% and 95% empirical quantiles of the variance estimates at points $t = 0.02, 0.04, \ldots, 2$. The best results for the Wald-type CIs are obtained with the second variance estimate $\hat{\sigma}^2_{2,nh}$ but the coverage proportions and average lengths do not outperform the results obtained with the bootstrap CIs in (3.2.9). Estimating the density g in $\hat{\sigma}_{1,nh}$ and $\hat{\sigma}_{2,nh}$ requires an additional bandwidth selection, whereas the estimate $\hat{\sigma}_{3,nh}$ is straightforward to obtain and does not suffer from a wrong bandwidth choice for g_{nh}. The variance of the first estimate $\hat{\sigma}^2_{1,nh}$ is larger than the variance of the second and third variance estimates $\hat{\sigma}^2_{2,nh}$ and $\hat{\sigma}^2_{3,nh}$, especially near the boundaries of the support.

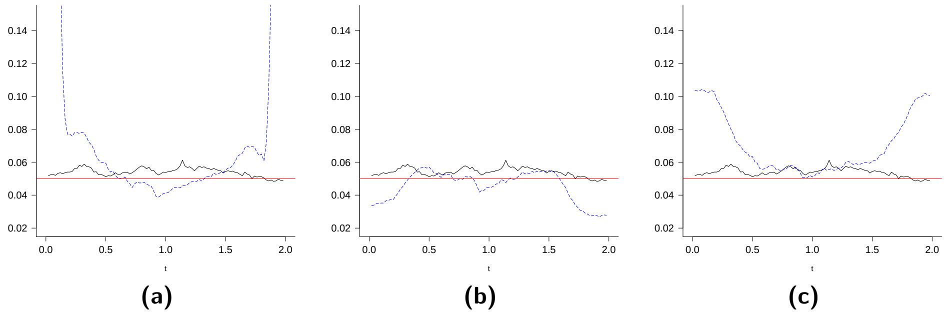


Figure 3.7: Uniform samples: Proportion of times that $F_0(t)$, $t = 0.02, 0.04, \ldots$ is not in the 95% CIs for the nonparametric bootstrap CIs defined in (3.2.9) (black, solid) and Wald-type CIs defined in (3.4.1) using the triweight kernel and (a) the first estimate $\hat{\sigma}^2_{1,nh}$ (blue, dashed), (b) the second estimate $\hat{\sigma}^2_{2,nh}$ (blue, dashed) and (c) the third estimate $\hat{\sigma}^2_{3,nh}$ (blue,dotted). $n = 1,000$, $N = 5,000$, $B = 1,000$ and $h = 2n^{-1/5}$.

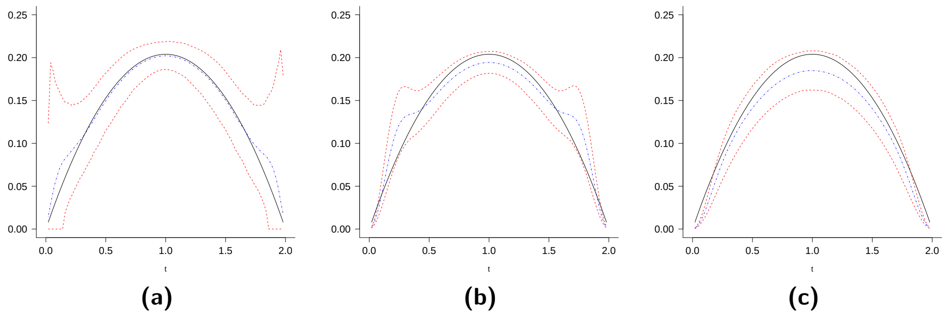


Figure 3.8: Uniform samples: True variance σ^2 (black, solid), mean estimate (blue, dashed-dotted) and the 5% and 95% empirical quantiles of the estimates (red, dashed) using the triweight kernel for (a) the first estimate $\hat{\sigma}^2_{1,nh}$, (b) the second estimate $\hat{\sigma}^2_{2,nh}$ and (c) the third estimate $\hat{\sigma}^2_{3,nh}$. $n = 1,000$, $N = 5,000$, $B = 1,000$ and $h = 2n^{-1/5}$.

3.4.2 Simulation study 2: Correcting the asymptotic bias

Although we have shown the validity of the bootstrap for constructing pointwise CIs around the SMLE, the performance of the CIs is often influenced by bias effects due to the fact that the SMLE is a biased estimate of the underlying distribution function. The MLE is an unbiased estimate of F_0, consequently, the MLE-based CIs do not suffer from bias issues. However, the results for the MLE-based CIs, shown in Figure 3.5 and Figure 3.6, indicate that these MLE-based intervals suffer from boundary effects and discreteness in the CIs.

To investigate the effect of the bias on the construction of the pointwise CIs in (3.2.9), we consider a second simulation study where the event times are generated from a truncated exponential distribution on $[0, 2]$ and the censoring times are uniformly distributed on $[0, 2]$. The density of the event times is given by $f_0(t) = \exp(-t)/(1 - \exp(-2))1_{[0,2]}(t)$ and therefore the bias β defined in (3.2.5) will influence the performance of the CIs. Since the bias issue is not due to the bootstrap algorithm used for constructing the CIs, we will show the results for the nonparametric bootstrap CIs proposed in Section 3.3.
Estimation of the bias is known to be a rather difficult task since it requires estimating the derivative f_0' of the density f_0 under current status data. Sufficiently accurate estimates of the bias are hard to obtain by direct estimation of f_0'. Besides estimating the derivative directly, we therefore also explore the effect of the bandwidth choice on the performance of the pointwise CIs. We first describe a procedure for selecting the bandwidth and next examine the quality of (a) a bootstrap based estimate of the bias, (b) a direct estimate of the bias using an estimate of f_0' and (c) undersmoothing the bandwidth, on the reduction of the bias effect present in the pointwise CIs. We first describe a data-driven bandwidth selection procedure.

3.4.2.1 Bandwidth selection

In the previous simulation study, the bandwidth was equal to $h = 2n^{-1/5}$, where the factor 2 was based on the size of the support $[0, 2]$ of the density f_0. This choice gave satisfactory results on the performance of the CIs discussed above. A bad choice of the bandwidth can however seriously affect the performance of the SMLE. It is therefore advisable to use an approach that selects the bandwidth with respect to some optimization criteria. We apply the method proposed in Hall (1990) to select the bandwidth which uses bootstrap subsamples of smaller size from the original sample to estimate the pointwise MSE of the SMLE. The method works as follows: To obtain an approximation to the optimal bandwidth minimizing the pointwise MSE, we generate B bootstrap subsamples of size $m = o(n)$ from the original sample using the subsampling principle and take $c_{t,opt}$ as the minimizer of

$$\widehat{MSE}(c) \stackrel{\text{def}}{=} B^{-1} \sum_{b=1}^{B} \left\{ \tilde{F}^b_{m,cm^{-1/5}}(t) - \tilde{F}_{nc_0n^{-1/5}}(t) \right\}^2, \tag{3.4.5}$$

where $\tilde{F}_{nc_0n^{-1/5}}$ is the SMLE in the original sample of size n using an initial bandwidth $c_0 n^{-1/5}$ for some constant c_0 and $\tilde{F}^b_{m,cm^{-1/5}}$ is the SMLE in the bth bootstrap run. The bandwidth used for estimating the SMLE is consequently given by $h = c_{t,opt} n^{-1/5}$. An important point is the fact that we have to use subsampling, i.e. bootstrapping with a smaller sample size, for estimating the right bandwidth in a reasonable fashion, as argued convincingly in Hall (1990). In the simulation study below we show the results for $m = 50$ when generating subsamples from a sample of size $n = 1,000$. Other subsample sizes $m = 30, 100$ were considered as well which resulted in similar optimal bandwidth choices (not shown here).

Figure 3.9 compares the proportion of times that $F_0(t)$ is not in the 95% bootstrap CIs for $t = 0.02, 0.04, \ldots, 2$ with the corresponding proportions in the bias corrected CIs given by

$$[\tilde{F}_{nh}(t) - Q^*_{1-\alpha/2}(t)\sqrt{S_{nh}(t)} - \beta(t)n^{-2/5},\\ \tilde{F}_{nh}(t) - Q^*_{\alpha/2}(t)\sqrt{S_{nh}(t)} - \beta(t)n^{-2/5}], \tag{3.4.6}$$

where $\beta(t)$ is the true bias of the SMLE at timepoint t defined in (3.2.5). The bandwidth of the SMLE is selected by the procedure described above. The coverage proportions of the uncorrected CIs are clearly smaller than the nominal 95%-level at the left endpoint of the interval $[0, 2]$ in correspondence to the region where $\beta(t)$ is largest and correcting for the bias effect is needed to obtain good CIs. Figure 3.9 suggests that the coverage proportions of the intervals will be satisfying if the bias can be estimated sufficiently accurate.

Estimation of the bias requires estimating the density f_0, which is a rather difficult task with current status data. A kernel based estimate of f_0' using the MLE F_n is given by

$$\tilde{f}'_{n\bar{h}}(t) \stackrel{\text{def}}{=} \bar{h}^{-2} \int K'\left((t-x)/\bar{h}\right)\, dF_n(x), \tag{3.4.7}$$

where the bandwidth $\bar{h} \sim n^{-1/9}$. In our experiments, we take the bandwidth of the estimate $\tilde{f}'_{n\bar{h}}(t)$ equal to $\bar{h} = \bar{c}_{t,opt}n^{-1/9}$ where $\bar{c}_{t,opt}$ is selected by the same bootstrap-MSE approach discussed in Section 3.4.2.1, but with the SMLE replaced by this derivative estimate. To obtain good estimates of f_0' near the boundaries of the support, we consider the boundary correction method explained in Section 9.2 of Groeneboom and Jongbloed (2014). A direct estimator of the actual bias is then obtained by first replacing $f_0'(t)$ in (3.2.5) by the estimate $\tilde{f}'_{nh}(t)$ and next multiplying with $n^{-2/5}$, i.e. the order of the actual bias that has to be taken into account when constructing the CIs.

Similarly to the estimate of the pointwise MSE defined in (3.4.5), we can also construct a bootstrap method for estimating the bias by using the subsampling principle described in Hall (1990). Our estimate $\widehat{Bias}(t)$ of the actual bias $\beta(t)n^{-2/5}$ is given by

$$\widehat{Bias}(t) \stackrel{\text{def}}{=} \left\{ B^{-1} \sum_{b=1}^{B} \left\{ \tilde{F}^b_{m,c_{t,opt}m^{-1/5}}(t) - \tilde{F}_{nc_0n^{-1/5}}(t) \right\} \right\} \left(\frac{m}{n}\right)^{2/5}.$$

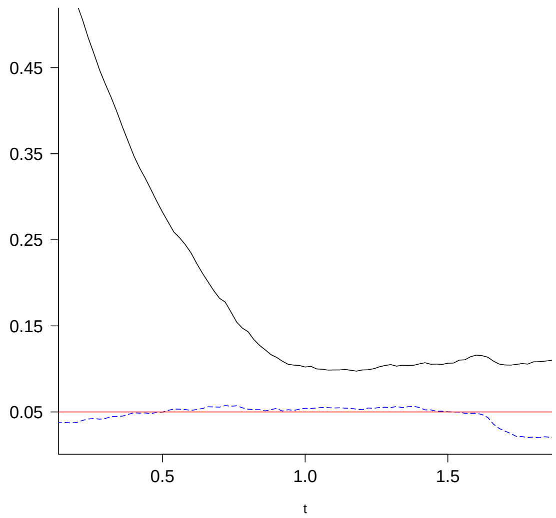


Figure 3.9: Truncated exponential samples: Proportion of times that $F_0(t)$, $t = 0.02, 0.04, \ldots$ is not in the 95% CIs for the nonparametric bootstrap CIs defined in (3.2.9) (black, solid) and the bias corrected CIs defined in (3.4.6) (blue, dashed). $n = 1,000$, $N = 5,000$, $B = 1,000, m = 50$ and $h = c_{t,opt}n^{-1/5}$.

Figure 3.10 compares the average true bias effect $\beta(t)n^{-2/5}$ and the average bias estimates obtained by either the direct estimation approach or the bootstrap based bias estimate for sample sizes $n = 1,000; 5,000$ and $n = 10,000$. Note that, since the bandwidth constant $c_{t,opt}$ used for estimating the SMLE is different in each simulation run, the true bias in each run is also different and therefore the average true bias is shown in Figure 3.10. The actual size of the bias decreases with increasing sample size and the results for the direct bias estimate using the estimate $\tilde{f}'_{n\bar{h}}$ are slightly better than the results for the bootstrap estimate of $\beta(t)n^{-2/5}$.

The proportion of times that $F_0(t)$ is not in the 95% bootstrap CIs, shown in Figure 3.11, decreases if one corrects for the bias by one of the discussed bias estimates. The coverage proportions are however still anti-conservative for points in the left end of the support. We also considered constructing the bias corrected CIs in the uniform model used in Section 3.4.1 where the actual bias is zero (results not shown). The results of the uncorrected CIs in (3.2.9) were slightly better and estimating the bias in this model has a somewhat negative effect on the coverage proportions of the pointwise CIs around the SMLE.

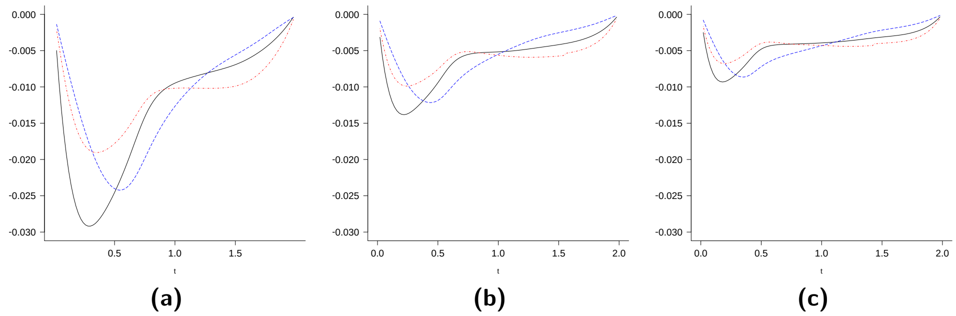


Figure 3.10: Truncated exponential samples: Average true bias (black solid) and average estimated bias for the bootstrap based estimate (blue, dashed) and the direct estimate (red, dashed-dotted) for samples (and subsamples) of size (a) $n = 1,000, m = 50$, (b) $n = 5,000, m = 100$ and (c) $n = 10,000, m = 250$. $N = 5,000$, $B = 1,000$ and $h = c_{t,opt}n^{-1/5}$.

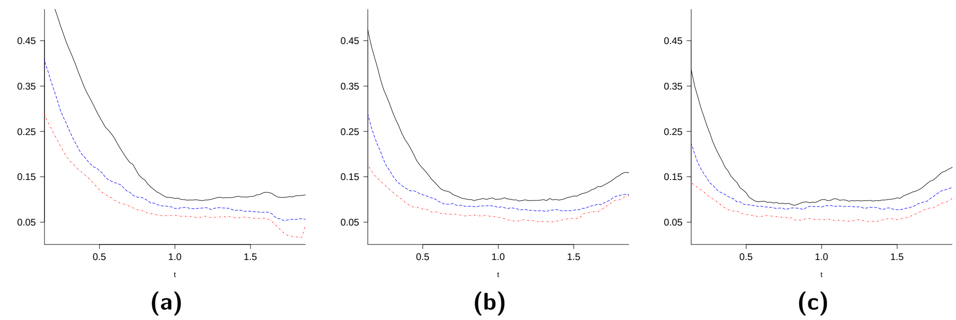


Figure 3.11: Truncated exponential samples: Proportion of times that $F_0(t)$, $t = 0.02, 0.04, \ldots$ is not in the 95% CIs defined in (3.2.9) (black, solid) and the bias corrected CIs defined in (3.4.6) with bootstrap based bias estimate (blue, dashed) and direct bias estimate (red, dashed-dotted) for samples (and subsamples) of size (a) $n = 1,000, m = 50$, (b) $n = 5,000, m = 100$ and (c) $n = 10,000, m = 250$. $N = 5,000$, $B = 1,000$ and $h = c_{t,opt}n^{-1/5}$.

We next investigate how the choice of the bandwidth can affect the coverage proportions and average length of our CIs. To this end, we consider the concept of undersmoothing proposed by Hall (1992) and take $c_{t,opt}n^{-1/4}$ as the bandwidth used in constructing the CIs defined in (3.2.9). The coverage proportions of the CIs for the exponential model, illustrated in Figure 3.12, show that the performance of the CIs around the SMLE improves by undersmoothing the bandwidth. We also observed that if we considered a smaller bandwidth choice $h = (1/3)c_{t,opt}n^{-1/5}$, the coverage proportions even improve further and give satisfactory results in the left end point of the support. This illustrates that a smaller bandwidth choice can indeed correct for the bias in the CIs. So far, we did not find any better bandwidth selection procedure than the one discussed in this paper and further research on bandwidth selection procedures is still needed. We also combined the technique of undersmoothing with direct bias estimation by constructing the CIs in (3.4.6) when using a bandwidth $c_{t,opt}n^{-1/4}$ and an estimate of $\beta(t)$. This resulted in slightly better CIs for the exponential model but gave worse results in the uniform model. The results of the CIs in (3.2.9) in the uniform model with a bandwidth $h = c_{t,opt}n^{-1/4}$ or $h = (1/3)c_{t,opt}n^{-1/5}$ are in line with the results obtained with a bandwidth $h = c_{t,opt}n^{-1/5}$ and similar to the results shown in Figure 3.7. This shows that undersmoothing the bandwidth in a model without bias has no negative effect on the coverage proportions of our CIs. By undersmoothing the bandwidth, the length of our SMLE-based CIs increases but the average length of the CIs remains remarkably smaller than the average length of the CIs around the MLE proposed by Banerjee and Wellner (2005) and Sen and Xu (2015) (see Table 3.2 for a comparison in simulation model 1 and 2).

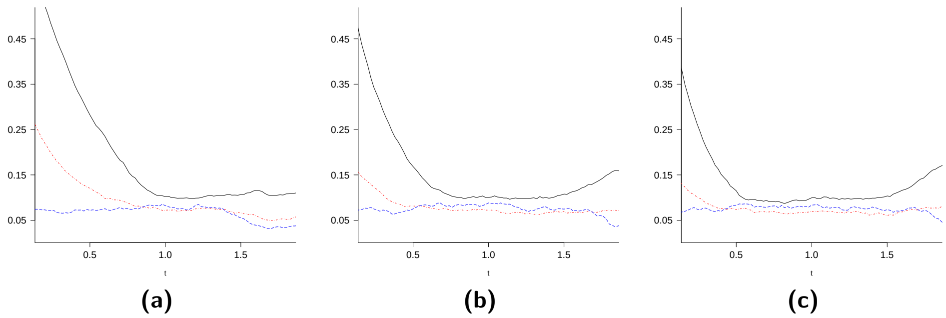


Figure 3.12: Proportion of times that $F_0(t)$, $t = 0.02, 0.04, \ldots$ is not in the 95% CIs defined in (3.2.9) with $h = c_{t,opt}n^{-1/5}$ (black, solid), $h = c_{t,opt}n^{-1/4}$ (red, dashed-dotted) and $h = (1/3)c_{t,opt}n^{-1/5}$ (blue, dashed) for samples (and subsamples) of size (a) $n = 1,000, m = 50$, (b) $n = 5,000, m = 100$ and (c) $n = 10,000, m = 250$. $N = 5,000$, $B = 1,000$.

3.5 Real data examples

3.5.1 Hepatitis A

Keiding (1991) considered a cross-sectional serological survey data on the presence of igG antibodies against Hepatitis A infection conducted in 1964 in Bulgaria. Samples were collected from school children and blood donors ($n = 850$), aged between 1 and 86 years old, and were tested for the presence or absence of such antibodies, thereby indicating

Table 3.2: Average length of the nonparametric SMLE-based CIs for different bandwidth choices (SMLE ($h \sim n^{-1/5}$) and SMLE ($h \sim n^{-1/4}$)) and average length of the MLE-based CIs proposed by Banerjee and Wellner (2005) (BW) and Sen and Xu (2015) (SX) at timepoints $t = 0.5, 1, 1.5$. $n = 1,000$ for samples from the uniform and truncated exponential distribution of simulation studies 1 and 2.

	Uniform			Exponential		
Method	$t = 0.5$	$t = 1$	$t = 1.5$	$t = 0.5$	$t = 1$	$t = 1.5$
SMLE ($h \sim n^{-1/5}$)	0.064819	0.077020	0.064976	0.085540	0.087565	0.057716
SMLE ($h \sim n^{-1/4}$)	0.079671	0.092096	0.079757	0.103828	0.101595	0.067480
MLE (BW)	0.164767	0.184590	0.165699	0.204079	0.161122	0.104002
MLE (SX)	0.183982	0.202430	0.186452	0.225882	0.176159	0.118541

past infection experience and leading to current status data. For Hepatitis A, lifelong humoral immunity after recovery from infection is presumed. The Hepatitis A virus is primarily transmitted via feco-oral contact and contact with blood products of infected individuals. Hepatitis A is a viral liver disease that can cause mild to severe illness upon contraction. The individual's age at the cross-sectional sampling is considered as the censoring time. We are interested in estimating the seroprevalence for Hepatitis A in Bulgaria. We constructed the Studentized SMLE-based smooth bootstrap CIs using a local bandwidth $h(t) = (0.5M + 1.5t)n^{-1/5}$, where $M = 86$ is the largest observed age. A picture of the CIs together with the LR-based CIs of Banerjee and Wellner (2005) and the CIs of Sen and Xu (2015) is given in Figure 3.13. The estimated prevalence of Hepatitis A at the age of 18 is 0.51, about half of the infections in Bulgaria happen during childhood. The length of the CIs is smallest for our SMLE-based CIs and largest for the Sen and Xu (2015) CIs. The latter CIs have left and right end points that are not monotone increasing in age, a property that is not shared by the other two CIs which have monotone increasing bounds. In contrast to the CIs of Banerjee and Wellner (2005), the bounds of our SMLE-based CIs and the CIs proposed by Sen and Xu (2015) are not increasing by construction. The applicability of the nonparametric bootstrap SMLE-based CIs on the Hepatitis A dataset can be found in the R-package `curstatCI`, where also the data-driven bandwidth procedure is illustrated.

3.5.2 Rubella

Keiding et al. (1996) considered a current status data set on the prevalence of rubella

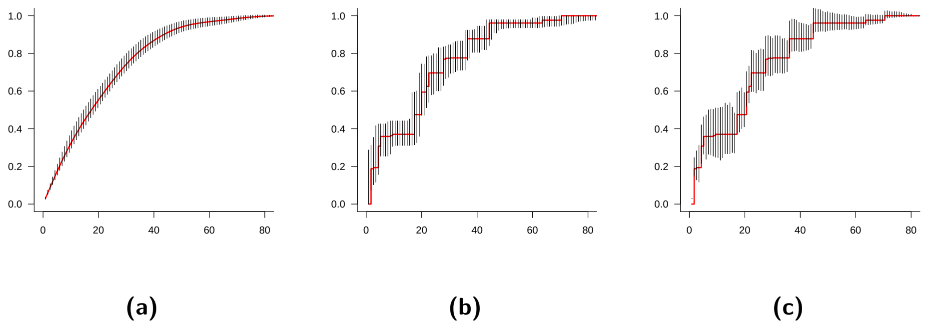


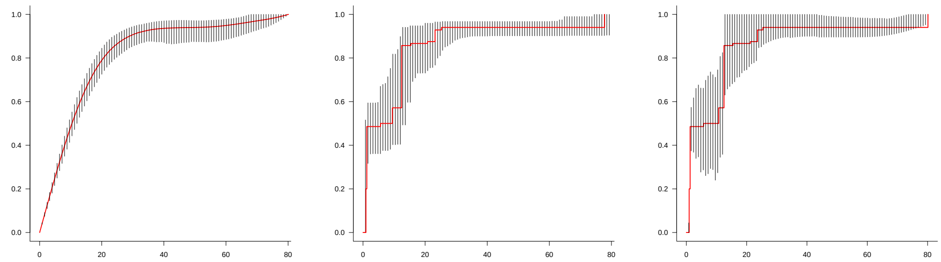
Figure 3.13: Hepatitis A data: (a) smooth bootstrap SMLE-based CIs defined in (3.2.9), (b) Banerjee and Wellner (2005) CIs and (c) Sen and Xu (2015) CIs based on $n = 850$ observations using $B = 1,000$ bootstrap samples. In (a) the SMLE (red, solid) is given and in (b,c) the MLE (red, step function) is given. $h(t) = (43 + 1.5t)n^{-1/5}$ for SMLE-based CIs and $h = 86n^{-1/5}$ for the CIs of Sen and Xu (2015).

in 230 Austrian males older than three months. Rubella is a highly contagious childhood disease spread by airborne and droplet transmission. The symptoms (such as rash, sore throat, mild fever and swollen glands) are less severe in children than in adults. Since the Austrian vaccination policy against rubella only vaccinated girls, the male individuals included in the dataset represent an unvaccinated population and (lifelong) immunity could only be acquired if the individual got the disease. We are interested in estimating the time to immunization (i.e. the time to infection) against rubella using the SMLE. Figure 3.14 shows the CIs obtained with the nonparametric bootstrap and illustrates the applicability of our method in a real data example.

3.6 Application: The current status linear regression model

As a consequence of Lemma 3.2.1 and Lemma 3.3.1, the bootstrap procedures of Section 3.2 and Section 3.3 can also be used to do inference in the current status linear regression model described in Chapter 2. In this section we use the nonparametric bootstrap procedure to construct CIs for the regression parameter $\boldsymbol{\alpha}_0$ in the current status linear regression model (2.1.1). Recall that the SSE of Section 2.4.1 is defined as a zero-crossing of

$$\sum_{\hat{F}_{n,\boldsymbol{\alpha}}(T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i) \in [\epsilon, 1-\epsilon]} \boldsymbol{X}_i \{ \Delta_i - \hat{F}_{n,\boldsymbol{\alpha}}(T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i) \}, \tag{3.6.1}$$



(a) **(b)** **(c)**

Figure 3.14: Rubella data: (a) nonparametric bootstrap SMLE-based CIs defined in (3.2.9), (b) Banerjee and Wellner (2005) CIs and (c) Sen and Xu (2015) CIs based on $n = 230$ observations using $B = 1,000$ bootstrap samples. In (a) the SMLE (red, solid) is given and in (b,c) the MLE (red, step function) is given. $h = c_{t,opt}n^{-1/4}$ ($c_{t,opt}$ is obtained by subsampling with $B = 1,000$ bootstrap samples of smaller size $m = 50$) for SMLE-based CIs and $h = 80n^{-1/5}$ for the CIs of Sen and Xu (2015).

where $\hat{F}_{n,\boldsymbol{\alpha}}$ is the MLE for fixed $\boldsymbol{\alpha}$ given in Section 2.3 and ϵ is a fixed truncation parameter in $(0, 1/2)$. A bootstrap version $\hat{\boldsymbol{\alpha}}_n^*$ based on a nonparemetric bootstrap sample from $\mathbb{P}_n$ is then defined as the zero-crossing of

$$\sum_{\hat{F}_{n,\boldsymbol{\alpha}}^*(T_i-\boldsymbol{\alpha}^T\boldsymbol{X}_i)\in[\epsilon,1-\epsilon]} M_{ni}\boldsymbol{X}_i\{\Delta_i - \hat{F}_{n,\boldsymbol{\alpha}}^*(T_i - \boldsymbol{\alpha}^T\boldsymbol{X}_i)\}, \tag{3.6.2}$$

where $\hat{F}_{n,\boldsymbol{\alpha}}^*$ is the MLE in the bootstrap sample and M_{ni} are the bootstrap weights. A straightforward extension of the results given in Section 3.3 shows that, as n tends to infinity,

$$\hat{E}_n \left| n^{-1/3}\left\{\hat{F}_{n,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) - F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\right\}\right|^p,$$

stays bounded in probability for all $(\boldsymbol{x}, t) \in \{(\boldsymbol{x}, t) : F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \in [\epsilon, 1-\epsilon]\}$ and for all $\boldsymbol{\alpha}$ in a neighborhood of $\boldsymbol{\alpha}_0$ where $F_{\boldsymbol{\alpha}}$ is defined by (2.3.1).
The validity of the bootstrap method follows from the fact that, in probability, we have conditionally on the data $(\boldsymbol{X}_1, T_1, \Delta_1), \ldots, (\boldsymbol{X}_n, T_n, \Delta_n)$ that,

$$\begin{aligned}-\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\alpha}}_n^* - \boldsymbol{\alpha}_n) = \sqrt{n}\int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} &\{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|T - \boldsymbol{\alpha}_0^T\boldsymbol{X} = t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\} \\ &\cdot\{\delta - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\}\, d(\hat{\mathbb{P}}_n - \mathbb{P}_n)(\boldsymbol{x}, t, \delta) \\ + o_{\hat{P}}(1 + \sqrt{n}\|\hat{\boldsymbol{\alpha}}_n^* - \boldsymbol{\alpha}_n\|),\end{aligned} \tag{3.6.3}$$

where the dominant term in the right-hand side of the display above is normally distributed with mean zero and variance $\boldsymbol{B}$ conditional on $(\boldsymbol{X}_1, T_1, \Delta_1), \ldots, (\boldsymbol{X}_n, T_n, \Delta_n)$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are defined in Theorem 2.4.1. The validity of the nonparametric bootstrap is further analyzed in Appendix B, Section B.2.3.

Remark 3.6.1. *The nonparametric bootstrap is also valid for the ESE of $\boldsymbol{\alpha}_0$ proposed in Section 2.4.2 based on a different score function involving the MLE $\hat{F}_{n,\boldsymbol{\alpha}}$ and the derivative of the SMLE $\tilde{F}_{nh,\boldsymbol{\alpha}}$. Moreover, by Lemma 3.2.1, the validity of the smooth bootstrap procedure of Section 3.2 can also be proved along the same lines.*

To provide more insight in the finite sample behavior of bootstrap estimators we show in Tables 3.3 and 3.4 the results of two simulation studies for a one-dimensional regression model $Y = \alpha_0 X + \varepsilon$. We only show the results for the nonparametric bootstrap estimates since the results for the smooth bootstrap estimates are similar. In the first simulation setting, we consider the simulation study given in Section 2.6 and we take $\alpha_0 = 0.5$ and consider Uniform(0,2) distributions for the variables T and X; for the density of the random error ε we take $f_0(e) = 384(e - 3/8)(5/8 - e)1_{[3/8,5/8]}(e)$. In the second simulation model, T, X and ε are independently sampled from a standard normal distribution and $\alpha_0 = 1$. A similar model was considered in Abrevaya (1999).

With these simulations we want to point out that it is not necessary to use smoothing techniques for doing inferences in the current status linear regression model. We compare the SSE with Han's maximum rank correlation estimator (Han (1987), MRCE) and with the ESE. The asymptotic behavior of the MRCE for the current status model, also obtained without any smoothing techniques, is established in Abrevaya (1999) where the author also proposes consistent kernel-based estimates of the asymptotic variance of the MRCE. We use these variance estimates to construct estimates for $\boldsymbol{A}, \boldsymbol{B}$ and the almost (determined by the truncation parameter ϵ) efficient variance of the ESE. For more details about the variance estimation we refer to Abrevaya (1999).

A summary of $N = 1,000$ simulation runs from models 1 and 2 for different sample sizes n is given in Tables 3.3 and 3.4. For each estimator, the mean, n times the variance and n times MSE is given in columns 3-5. The asymptotic variance of the estimators equals 0.193612 for the SSE, 0.158699 for the ESE and 0.192857 for the MRCE in model 1 using truncation parameter $\epsilon = 0.001$. The corresponding asymptotic variances in model 2 are equal to 5.046413, 4.994988 and 5.35448 respectively. The asymptotic variance of the SSE without truncation (i.e. $\epsilon = 0$) equals the asymptotic variance of the MRCE in model 1. The efficient variances are 0.151706 in model 1 and 4.994987 in model 2. Note that the differences between the limiting variances for the different estimation methods are tiny and that the effect of the truncation parameter ϵ on the asymptotic behavior of the score estimators is small. Tables 3.3 and 3.4 show that n times the variance tends to converge to the asymptotic variance for all estimators. The ESE performs worse for small sample sizes and the results suggest to use the SSE for point estimation of the regression parameter α_0. Also note that the results in Table 3.3 are slightly different than the results in Table 2.1 due to the fact that we redid the simulation study, thereby sampling new data sets from the underlying simulation model.

We constructed Wald-type CIs, similar to the intervals proposed in Abrevaya (1999), using the asymptotic normal limiting distribution of the estimators and compared the coverage proportion and average length of these intervals with bootstrap CIs based on the nonparametric bootstrap described in this chapter using $B = 1,000$ samples from the original data. For the MRCE, the validity of the nonparametric bootstrap is proved in Subbotin (2007). The Wald-type CIs remain anti-conservative for the ESE in model 2.

We observed (result not shown) that, in both models, the bias in estimating the efficient variance of the ESE remains larger than the bias of the asymptotic variance estimates for the SSE and the MRCE. Tables 3.3 and 3.4 show that the coverage proportion of the nonparametric bootstrap CIs converges to the nominal $95\%-$level. The average length of the CIs obtained by resampling from the original data is smaller than the corresponding length of the Wald-type CIs for all methods in the first simulation study and for the SSE in the second simulation study. We also investigated the behavior of Studentized bootstrap CIs (results not shown) based on the variance estimate used in the construction of the Wald-type CIs, but no improvement was observed for the behavior of the bootstrap intervals.

Our results do not indicate better performances corresponding to smoothing techniques and therefore suggest that smoothing should not be the primary concern in inferences for the current status linear regression model. Note that the Wald-type CIs are constructed using smoothing kernel estimation for the variance estimate and that the only results obtained without any smoothing are the bootstrap CIs for the SSE and the MRCE. It is noteworthy that the SSE tends to perform better than the MRCE, which is not based on a nuisance parameter that is not estimable at $\sqrt{n}-$rate. Based on these results, we recommend the use of the SSE in combination with the nonparametric bootstrap procedure for doing inference in the current status linear regression model.

Table 3.3: Simulation model 1: The mean value, n times the variance and n times MSE of the simple score estimate (SSE), the maximum rank correlation estimate (MRCE) and the efficient score estimate (ESE). CP: coverage proportion of 95% CIs (Wald-type intervals based on a kernel variance estimate and nonparametric bootstrap intervals) that contain the true parameter value $\alpha_0 = 0.5$, AL: Average length of the CIs for different samples sizes n based on $N = 1,000$ simulation runs and $B = 1,000$ bootstrap samples. $\epsilon = 0.001$.

Estimate	n	mean	$n\times$var	$n\times$MSE	Wald-type CI		Bootstrap CI	
					CP	AL	CP	AL
SSE	100	0.498943	0.310723	0.310968	0.978	0.265883	0.824	0.204163
	500	0.499717	0.220885	0.220925	0.982	0.097457	0.897	0.080317
	1,000	0.500720	0.217415	0.217933	0.977	0.065837	0.924	0.055648
	5,000	0.499993	0.195111	0.195112	0.977	0.027159	0.945	0.024423
MRCE	100	0.497996	0.308180	0.308582	0.979	0.268731	0.821	0.205522
	500	0.499761	0.251232	0.251260	0.978	0.098028	0.862	0.089143
	1,000	0.500553	0.246388	0.246693	0.973	0.063990	0.911	0.053129
	5,000	0.499876	0.208386	0.208462	0.965	0.027197	0.922	0.026987
ESE	100	0.500145	0.337755	0.337757	0.964	0.252687	0.824	0.223849
	500	0.499671	0.217428	0.217482	0.978	0.094390	0.896	0.080003
	1,000	0.500742	0.207401	0.207953	0.973	0.063990	0.911	0.053129
	5,000	0.500228	0.185614	0.185874	0.972	0.026396	0.904	0.022285

Table 3.4: Simulation model 2: The mean value, n times the variance and n times MSE of the simple score estimate (SSE), the maximum rank correlation estimate (MRCE) and the efficient score estimate (ESE). CP: coverage proportion of 95% CIs (Wald-type intervals based on a kernel variance estimate and nonparametric bootstrap intervals) that contain the true parameter value $\alpha_0 = 1$, AL: Average length of the CIs, for different samples sizes n based on $N = 1,000$ simulation runs and $B = 1,000$ bootstrap samples. $\epsilon = 0.001$.

Estimate	n	mean	$n\times$var	$n\times$MSE	Wald-type CI		Bootstrap CI	
					CP	AL	CP	AL
SSE	100	0.935732	4.525330	4.938096	0.922	1.000283	0.855	0.79952
	500	0.966217	4.676249	5.246881	0.926	0.399728	0.902	0.364210
	1,000	0.977799	5.032432	5.525339	0.933	0.279928	0.914	0.262449
	5,000	0.989466	4.580756	5.135616	0.945	0.124375	0.948	0.121388
MRCE	100	1.038510	8.500588	8.648890	0.925	1.125225	0.889	1.364034
	500	1.006050	6.443404	6.461690	0.932	0.429007	0.912	0.473787
	1,000	1.002680	6.294143	6.301326	0.939	0.296537	0.903	0.320908
	5,000	0.998502	5.160694	5.171915	0.962	0.129512	0.954	0.136487
ESE	100	0.974199	5.722576	5.789144	0.768	0.604649	0.827	0.910229
	500	0.998806	5.984291	5.985003	0.823	0.290297	0.902	0.430819
	1,000	1.005545	6.032743	6.063495	0.841	0.214280	0.928	0.302124
	5,000	1.002462	5.244373	5.274692	0.892	0.104281	0.951	0.131427

Chapter 4

Single index models

Abstract

We develop an estimation technique for the regression parameter in the single index model, given by $\mathbb{E}(Y|\boldsymbol{X}) = \psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})$ where the link function ψ_0 is monotone increasing and Y is the response and $\boldsymbol{X}$ are the covariates. For this semiparametric model it has been proposed to estimate the regression parameter $\boldsymbol{\alpha}_0$ via the profiled least squares method where first the link function ψ_0 is estimated nonparametrically by the monotone least square estimate (LSE) $\hat{\psi}_{n,\boldsymbol{\alpha}}$ for each $\boldsymbol{\alpha}$ and next the estimate of the regression parameter is obtained by minimizing the sum of squared deviations $\sum_{i=1}^n \{Y_i - \hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\}^2$ over $\boldsymbol{\alpha}$. Although it is natural to propose this least squares procedure, it is still unknown whether it will produce $\sqrt{n}$-consistent estimates. We show that the latter property will hold if we solve a score equation corresponding to this minimization problem. This is the first time that $\sqrt{n}$-consistent estimates are constructed based on the piecewise constant LSE of the link function. Our simulation studies do not give conclusive answers on the behavior of the profiled LSE but show that even if the profiled LSE of the regression parameter leads at all to a $\sqrt{n}$-consistent estimate, its performance is certainly inferior to the score procedures we propose in this Chapter. Good performances of our estimation approach, both asymptotically and numerically, are illustrated by the asymptotic normality of our score estimates and by simulation studies that show comparable or even better behavior of our score estimates compared to the rank estimates proposed by Han (1987) and Cavanagh and Sherman (1998).

4.1 Model description

Consider the following regression model

$$Y = \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X}) + \varepsilon, \tag{4.1.1}$$

where Y is a one-dimensional random variable, $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is a d-dimensional random vector with distribution G and ε is a one-dimensional random variable such that $\mathbb{E}[\varepsilon|\boldsymbol{X}] = 0$ G-almost surely. The function ψ_0 is a monotone link function in $\mathcal{M}$, where $\mathcal{M}$ is the set of monotone increasing functions defined on $\mathbb{R}$ and $\boldsymbol{\alpha}_0$ is a vector of regression parameters belonging to the $d-1$ dimensional sphere $\mathcal{S}_{d-1} = \{\boldsymbol{\alpha} \in \mathbb{R}^d : \|\boldsymbol{\alpha}\| = 1\}$.

4.2 Behavior of the least squares estimator of the link function

Let $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ denote n random variables which are i.i.d. like $(\boldsymbol{X}, Y)$ in (4.1.1), i.e. $\mathbb{E}(Y|\boldsymbol{X}) = \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X})$ G-almost surely and consider the sum of squared errors

$$S_n(\boldsymbol{\alpha}, \psi) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \Big\{ Y_i - \psi(\boldsymbol{\alpha}^T \boldsymbol{X}_i) \Big\}^2, \tag{4.2.1}$$

which can be computed for any pair $(\boldsymbol{\alpha}, \psi) \in \mathcal{S}_{d-1} \times \mathcal{M}$. The LSE $(\hat{\boldsymbol{\alpha}}_n, \hat{\psi}_n)$ is defined by

$$(\hat{\boldsymbol{\alpha}}_n, \hat{\psi}_n) \stackrel{\text{def}}{=} \arg \min_{\boldsymbol{\alpha} \in \mathcal{S}_{d-1}, \psi \in \mathcal{M}} S_n(\boldsymbol{\alpha}, \psi). \tag{4.2.2}$$

The LSE can be obtained as follows. For a fixed $\boldsymbol{\alpha} \in \mathcal{S}_{d-1}$, order the values $\boldsymbol{\alpha}^T \boldsymbol{X}_1, \ldots, \boldsymbol{\alpha}^T \boldsymbol{X}_n$ in increasing order and arrange $Y_1, \ldots, Y_n$ accordingly. As ties are not excluded, let $m = m_{\boldsymbol{\alpha}}$ be the number of distinct projections among $\boldsymbol{\alpha}^T \boldsymbol{X}_i$ and $Z_1^{\boldsymbol{\alpha}} < \ldots < Z_m^{\boldsymbol{\alpha}}$ the corresponding ordered values. For $1 \le i \le m$, let

$$n_i^{\boldsymbol{\alpha}} = \sum_{j=1}^{n} 1_{\{\boldsymbol{\alpha}^T \boldsymbol{X}_j = Z_i^{\boldsymbol{\alpha}}\}} \quad \text{and} \quad Y_i^{\boldsymbol{\alpha}} = \sum_{j=1}^{n} Y_j 1_{\{\boldsymbol{\alpha}^T \boldsymbol{X}_j = Z_i^{\boldsymbol{\alpha}}\}} / n_i^{\boldsymbol{\alpha}}.$$

Then, well-known results from isotonic regression theory imply that the functional $\psi \mapsto S_n(\boldsymbol{\alpha}, \psi)$ is minimized by the left derivative of the greatest convex minorant of the cumulative sum diagram

$$\Big\{ (0,0), \Big(\sum_{j=1}^{i} n_j^{\boldsymbol{\alpha}}, \sum_{j=1}^{i} n_j^{\boldsymbol{\alpha}} Y_j^{\boldsymbol{\alpha}} \Big), 1 \le i \le m \Big\}.$$

See for example Theorem 1.1 in Barlow et al. (1972) or Theorem 1.2.1 in Robertson et al. (1988). By strict convexity of $\psi \mapsto S_n(\psi, \boldsymbol{\alpha})$, the minimizer is unique at the

distinct projections. We denote by $\hat{\psi}_{n,\boldsymbol{\alpha}}$ the monotone function which takes the values of this minimizer at the distinct projections and is a stepwise and right-continuous function outside the set of those projections.
We first list below the assumptions needed to prove the asymptotic results stated in the remainder of the chapter.

Assumptions A1-A6

A1. The space $\mathcal{X}$ is convex, with a nonempty interior. There exists also $R > 0$ such that $\mathcal{X} \subset \mathcal{B}(0, R)$.

A2. There exists $K_0 > 0$ such that the true link function ψ_0 satisfies $|\psi_0(u)| \leq K_0$ for all u in $\{\boldsymbol{\alpha}^T\boldsymbol{x}, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\alpha} \in \mathcal{S}_{d-1}\}$.

A3. There exists $\delta_0 > 0$ such that the function $u \mapsto \mathbb{E}\big[\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = u\big]$ is monotone increasing on $\mathcal{I}_{\boldsymbol{\alpha}} = \{\boldsymbol{\alpha}^T\boldsymbol{x}, \boldsymbol{x} \in \mathcal{X}\}$ for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0) = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\| \leq \delta_0\}$.

A4. Let a_0 and b_0 denote the infimum and supremum of the interval $\mathcal{I}_{\boldsymbol{\alpha}_0} = \big\{\boldsymbol{\alpha}_0^T\boldsymbol{x},\ \boldsymbol{x} \in \mathcal{X}\big\}$. Then, the true link function ψ_0 is continuously differentiable on $(a_0 - \delta_0 R, b_0 + \delta_0 R)$, where R is the same radius of assumption A1 above, and there exists $C > 0$ such that $\psi_0' \geq C$ on $(a_0 - \delta_0 R, b_0 + \delta_0 R)$.

A5. The distribution of $\boldsymbol{X}$ admits a density g, which is differentiable on $\mathcal{X}$. Also, there exist positive constants $\underline{c}_0$, $\bar{c}_0$, $\underline{c}_1$ and $\bar{c}_1$ such that $\underline{c}_0 \leq g \leq \bar{c}_0$ and $\underline{c}_1 \leq \partial g/\partial x_i \leq \bar{c}_1$ on $\mathcal{X}$ for all $1 \leq i \leq d$.

A6. There exist $a_0, M_0 > 0$ such that $\mathbb{E}\Big[|Y|^m\ |\boldsymbol{X} = \boldsymbol{x}\Big] \leq m! M_0^{m-2} a_0$ for all integers $m \geq 2$ and $\boldsymbol{x} \in \mathcal{X}$ G-almost surely.

Assumption A1 ensures that the support of the linear predictor $\boldsymbol{\alpha}^T\boldsymbol{X}$ is an interval for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$. Assumption A3 is made to enable deriving the explicit limit of the LSE $\hat{\psi}_{n,\boldsymbol{\alpha}}$ for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$. In Lemma C.5.6, given in Appendix C, Section C.5, we will show the plausibility of this Assumption A3 by proving that for $\boldsymbol{\alpha}$ in a neighborhood of $\boldsymbol{\alpha}_0$ the derivative of the function $u \mapsto \mathbb{E}\big[\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = u\big]$ is indeed strictly positive if the derivative of the true link function stays away from zero. In order to prove this result, we hence need Assumption A4 on the positiveness of the derivative ψ_0'. Assumption A6 is needed to show that $\max_{1\leq i\leq n}|Y_i| = O_p(\log n)$. As noted in Balabdaoui et al. (2016), such an assumption is satisfied if the conditional distribution of $Y|\boldsymbol{X} = \boldsymbol{x}$ belongs to an exponential family.

In Figure 4.1 we compare the true link function ψ_0 with the function $u \mapsto \mathbb{E}\big[\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = u\big]$ for the model $E(Y|\boldsymbol{X}) = \psi_0(\alpha_{01}X_1 + \alpha_{02}X_1)$, where $X_1, X_2 \overset{i.i.d}{\sim} U[0,1]$, $\psi_0(\boldsymbol{x}) = x^3$ and $\alpha_{01} = \alpha_{02} = 1/\sqrt{2}$ for $\alpha_1 = 1/2, \alpha_2 = \sqrt{3}/2$. Figure 4.1 illustrates the monotonicity of the function introduced in Assumption A3.

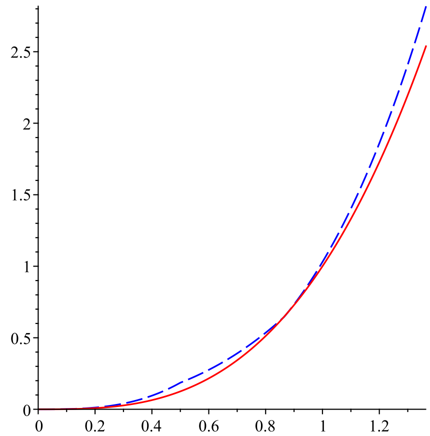


Figure 4.1: The real ψ_0 (red, solid) and the function $u \mapsto \mathbb{E}\big[\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = u\big]$ (blue, dashed) for $\psi_0(x) = x^3$, $\alpha_{01} = \alpha_{02} = 1/\sqrt{2}$ and $\alpha_1 = 1/2, \alpha_2 = \sqrt{3}/2$, with $X_1, X_2 \overset{i.i.d}{\sim} U[0,1]$.

We have the following results.

Proposition 4.2.1. *Suppose that Assumptions A1-A3 hold and let the function $\psi_{\boldsymbol{\alpha}}$ be defined by*

$$\psi_{\boldsymbol{\alpha}}(u) \overset{\text{def}}{=} \mathbb{E}\big[\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = u\big]. \tag{4.2.3}$$

Then, the functional $L_{\boldsymbol{\alpha}}$ given by,

$$\psi \mapsto L_{\boldsymbol{\alpha}}(\psi) = \int_{\mathcal{X}} \Big(\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{x}) - \psi(\boldsymbol{\alpha}^T\boldsymbol{x})\Big)^2 dG(\boldsymbol{x}), \tag{4.2.4}$$

admits a minimizer $\hat{\psi}^{\boldsymbol{\alpha}}$, over the set of monotone increasing functions defined on $\mathbb{R}$, denoted by $\mathcal{M}$, such that $\hat{\psi}^{\boldsymbol{\alpha}}$ is uniquely given by the function $\psi_{\boldsymbol{\alpha}}$ in (4.2.3) on $\mathcal{I}_{\boldsymbol{\alpha}} = \Big\{\boldsymbol{\alpha}^T\boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}\Big\}$.

Proposition 4.2.2. *Under Assumptions A1-A6, we have,*

$$\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} \int \Big\{\hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\Big\}^2 dG(\boldsymbol{x}) = O_p\Big((\log n)^2 n^{-2/3}\Big).$$

The proofs of Proposition 4.2.1 and Proposition 4.2.2 are given in Appendix C.

4.3 $\sqrt{n}$-consistent regression parameter estimation on the unit sphere

4.3.1 The simple score estimator

Consider the problem of minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-\hat\psi_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\right\}^2, \tag{4.3.1}$$

over all $\boldsymbol{\alpha}\in\mathcal{S}_{d-1}$, where $\hat\psi_{n,\boldsymbol{\alpha}}$ is the LSE of $\psi_{\boldsymbol{\alpha}}$. Since the parameter space is $\mathcal{S}_{d-1}$, we need to estimate the regression parameter $\boldsymbol{\alpha}_0$ in a $d-1$ dimensional subspace of $\mathbb{R}^d$. We therefore introduce a new parameter vector in $\mathbb{R}^{d-1}$ via a local parametrization mapping $\mathbb{R}^{d-1}$ to the sphere $\mathcal{S}_{d-1}$. For each $\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)$ on the sphere $\mathcal{S}_{d-1}$, there exists a unique vector $\boldsymbol{\beta}\in\mathbb{R}^{d-1}$ such that

$$\boldsymbol{\alpha}=\mathbb{S}\left(\boldsymbol{\beta}\right).$$

We now construct an estimation algorithm to estimate the parameter $\boldsymbol{\beta}_0$ defined by $\boldsymbol{\alpha}_0=\mathbb{S}\left(\boldsymbol{\beta}_0\right)$ and obtain the final estimate of $\boldsymbol{\alpha}_0$ after applying the parametrization $\mathbb{S}$ to the estimate for $\boldsymbol{\beta}_0$. The minimization problem given in (4.3.1) is equivalent to minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-\hat\psi_{n,\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}_i\right)\right\}^2, \tag{4.3.2}$$

over all $\boldsymbol{\beta}$ where $\hat\psi_{n,\boldsymbol{\alpha}}$ is the LSE of $\psi_{\boldsymbol{\alpha}}$ with $\boldsymbol{\alpha}=\mathbb{S}(\boldsymbol{\beta})$. Analogously to the treatment of the score approach in the current status regression model proposed in Chapter 2, we consider the derivative of (4.3.2) w.r.t. $\boldsymbol{\beta}$, where we ignore the non-differentiability of the LSE $\hat\psi_{n,\boldsymbol{\alpha}}$. This leads to the set of equations,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T\boldsymbol{X}_i\left\{Y_i-\hat\psi_{n,\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}_i\right)\right\}=\mathbf{0}, \tag{4.3.3}$$

where $\boldsymbol{J}_{\mathbb{S}}$ is the Jacobian of the map $\mathbb{S}$ and where $\mathbf{0}\in\mathbb{R}^{d-1}$ is the vector of zeros. Just as in the analogous case of the simple score equation in Chapter 2, we cannot hope to solve equation (4.3.3) exactly. Instead, we define the solution in terms of a "zero-crossing" of the above equation where a zero-crossing is defined in Definition 2.4.1.
Our index score estimator $\hat{\boldsymbol{\alpha}}_n$ is defined by,

$$\hat{\boldsymbol{\alpha}}_n\stackrel{\text{def}}{=}\mathbb{S}(\hat{\boldsymbol{\beta}}_n), \tag{4.3.4}$$

where $\hat{\boldsymbol{\beta}}_n$ is a zero crossing of the function

$$\xi_{1,n}(\boldsymbol{\beta})\stackrel{\text{def}}{=}\int\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T\boldsymbol{x}\left\{y-\hat\psi_{n,\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\}\,d\mathbb{P}_n(\boldsymbol{x},y), \tag{4.3.5}$$

and $\mathbb{P}_n$ denotes the empirical probability measure of $(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)$. The probability measure of $(\boldsymbol{X}, Y)$ will be denoted by P_0 in the remainder of the chapter.
In addition to Assumptions A1-A6 above, the following assumptions will also be made.

Assumptions A7-A9

A7. For all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ such that $\mathbb{S}(\boldsymbol{\beta}) \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$, the random variable

$$\text{Cov}\Big[(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^T \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})^T \boldsymbol{X}, \psi_0\big(\mathbb{S}(\boldsymbol{\beta}_0)^T \boldsymbol{X}\big) \mid \mathbb{S}(\boldsymbol{\beta})^T \boldsymbol{X}\Big],$$

is not equal to 0 almost surely.

A8. The functions $\boldsymbol{J}_{\mathbb{S}}^{ij}(\boldsymbol{\beta})$, where $\boldsymbol{J}_{\mathbb{S}}^{ij}(\boldsymbol{\beta})$ denotes the $i \times j$ entry of $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})$ for $1 \le i \le d$ and $1 \le j \le d-1$ are $d-1$ times continuously differentiable on $\mathcal{C} = \{\boldsymbol{\beta} \in \mathbb{R}^{d-1} : \mathbb{S}(\boldsymbol{\beta}) \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)\}$ and there exists $M > 0$ satisfying

$$\max_{k. \le d-1} \sup_{\boldsymbol{\beta} \in \mathcal{C}} |D^k \boldsymbol{J}_{\mathbb{S}}^{ij}(\boldsymbol{\beta})| \le M, \tag{4.3.6}$$

where $k = (k_1, \dots, k_d)$ with k_j an integer $\in \{0, \dots, d-1\}$, $k. = \sum_{i=1}^{d-1} k_i$ and

$$D^k s(\boldsymbol{\beta}) \equiv \frac{\partial^{k.} s(\boldsymbol{\beta})}{\partial \beta_{k_1} \dots \partial \beta_{k_d}}.$$

We also assume that $\mathcal{C}$ is a convex and bounded set in $\mathbb{R}^{d-1}$ with a nonempty interior.

A9. $(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \, \mathbb{E}\Big[\psi_0'(\boldsymbol{\alpha}_0^T \boldsymbol{X}) \, \text{Cov}(\boldsymbol{X} | \boldsymbol{\alpha}_0^T \boldsymbol{X})\Big] \, (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))$ is nonsingular.

Theorem 4.3.1. *Let Assumptions A1-A9 be satisfied. Let also $\hat{\boldsymbol{\alpha}}_n$ be defined by (4.3.4). Then*

(i) [Existence of a root] A crossing of zero $\hat{\boldsymbol{\beta}}_n$ of $\xi_{1,n}(\boldsymbol{\beta})$ exists with probability tending to one.

(ii) [Consistency]

$$\hat{\boldsymbol{\alpha}}_n \xrightarrow{p} \boldsymbol{\alpha}_0, \qquad n \to \infty.$$

(iii) [Asymptotic normality] Define the matrices,

$$\boldsymbol{A} \stackrel{\text{def}}{=} \mathbb{E}\Big[\psi_0'(\boldsymbol{\alpha}_0^T \boldsymbol{X}) \, \textit{Cov}(\boldsymbol{X} | \boldsymbol{\alpha}_0^T \boldsymbol{X})\Big], \tag{4.3.7}$$

and

$$\boldsymbol{\Sigma} \stackrel{\text{def}}{=} \mathbb{E}\left[\left\{Y - \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X})\right\}^2 \left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X} | \boldsymbol{\alpha}_0^T \boldsymbol{X})\right\} \left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X} | \boldsymbol{\alpha}_0^T \boldsymbol{X})\right\}^T\right]. \tag{4.3.8}$$

Then

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) \xrightarrow{d} N_d\left(\mathbf{0}, \boldsymbol{A}^{-}\boldsymbol{\Sigma}\boldsymbol{A}^{-}\right),$$

where $\boldsymbol{A}^{-}$ is the Moore-Penrose inverse of $\boldsymbol{A}$.

Remark 4.3.1. *Note that $\boldsymbol{\alpha}_0^T\boldsymbol{A} = \mathbf{0}$ and that the normal distribution $N_d\left(\mathbf{0}, \boldsymbol{A}^{-}\boldsymbol{\Sigma}\boldsymbol{A}^{-}\right)$ is concentrated on the $(d-1)$-dimensional subspace, orthogonal to $\boldsymbol{\alpha}_0$ and is therefore degenerate, as is also clear from its covariance matrix $\boldsymbol{A}^{-}\boldsymbol{\Sigma}\boldsymbol{A}^{-}$, which is a matrix of rank $d-1$.*

To obtain the asymptotic normality result of the score estimator $\hat{\boldsymbol{\alpha}}_n$ given in Theorem 4.3.1, we prove in Appendix C the following asymptotic relationship for $\hat{\boldsymbol{\beta}}_n$:

$$\begin{aligned}
&\boldsymbol{B}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right)\\
&= \int \left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T \left\{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X} = \mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y - \psi_0\big(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\big)\right\}\\
&\qquad\qquad\qquad\qquad d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x}, y)\\
&\quad + o_p\left(n^{-1/2} + \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|\right).
\end{aligned}$$

where

$$\boldsymbol{B} = \left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T \mathbb{E}\Big[\psi_0'(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X})\,\mathsf{Cov}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X})\Big]\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right) = \left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T \boldsymbol{A}\,\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0), \tag{4.3.9}$$

in $\mathbb{R}^{(d-1)\times(d-1)}$. We assume in Assumption A9 that $\boldsymbol{B}$ is invertible so that

$$\begin{aligned}
&\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right)\\
&= \sqrt{n}\boldsymbol{B}^{-1}\int \left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T \left\{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X} = \mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y - \psi_0\big(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\big)\right\}\\
&\qquad\qquad\qquad\qquad d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x}, y)\\
&\quad + o_p\left(1 + \sqrt{n}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|\right)\\
&\xrightarrow{d} N_{d-1}(\mathbf{0}, \boldsymbol{\Pi}),
\end{aligned}$$

where

$$\boldsymbol{\Pi} = \boldsymbol{B}^{-1}\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T \boldsymbol{\Sigma}\,\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\,\boldsymbol{B}^{-1} \in \mathbb{R}^{(d-1)\times(d-1)}. \tag{4.3.10}$$

The limit distribution of the single index score estimator $\hat{\boldsymbol{\alpha}}_n$ defined in (4.3.4) now follows by an application of the delta method and we conclude that

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) &= \sqrt{n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n) - \mathbb{S}(\boldsymbol{\beta}_0)\right) = J_{\mathbb{S}}(\boldsymbol{\beta}_0)\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_p\left(\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right)\\
&\xrightarrow{d} N_d\left(\mathbf{0}, J_{\mathbb{S}}(\boldsymbol{\beta}_0)\boldsymbol{\Pi}\left(J_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T\right) = N_d\left(\mathbf{0}, \boldsymbol{A}^{-}\boldsymbol{\Sigma}\boldsymbol{A}^{-}\right),
\end{aligned}$$

where the last equality follows from the following lemma.

Lemma 4.3.1. *Let the matrix* $\boldsymbol{A}$ *be defined by (4.3.7) and let* $\boldsymbol{A}^-$ *be the Moore-Penrose inverse of* $\boldsymbol{A}$. *Then*

$$\boldsymbol{A}^- = \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T = \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\boldsymbol{B}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T.$$

The proof of Lemma 4.3.1 is given in Appendix C. An example of the mapping $\mathbb{S}$ and corresponding matrix $\boldsymbol{J}_{\mathbb{S}}$ is given in Section 4.4.

Remark 4.3.2. *For each map* $\mathbb{S}$ *and each parameter vector* $\boldsymbol{\beta}$, *we have*

$$(\mathbb{S}(\boldsymbol{\beta}))^T\mathbb{S}(\boldsymbol{\beta}) = 1.$$

Taking derivatives w.r.t. $\boldsymbol{\beta}$, *we get*

$$(\mathbb{S}(\boldsymbol{\beta}))^T\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}) = \boldsymbol{0}^T,$$

so that the columns of $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})$ *belong to the space*

$$\{\boldsymbol{\alpha}\}^\perp \equiv \{\mathbb{S}(\boldsymbol{\beta})\}^\perp \equiv \left\{\boldsymbol{z}\in\mathbb{R}^d : \boldsymbol{\alpha}^T z = 0\right\} \equiv \left\{\boldsymbol{z}\in\mathbb{R}^d : (\mathbb{S}(\boldsymbol{\beta}))^T z = 0\right\}.$$

By Lemma 4.3.1 it is now easy to see that also $\boldsymbol{\alpha}_0^T\boldsymbol{A}^- = \boldsymbol{0}$. *It is shown in Lemma 1 of Kuchibhotla and Patra (2017) that it is possible to construct a set of "local parametrization matrices"* $H_{\boldsymbol{\alpha}}$ *for each* $\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)$ *with* $\|\boldsymbol{\alpha}\| = 1$ *satisfying*

$$\boldsymbol{\alpha}^T H_{\boldsymbol{\alpha}} = \boldsymbol{0}^T \quad \text{and} \quad (H_{\boldsymbol{\alpha}})^T H_{\boldsymbol{\alpha}} = \boldsymbol{I}_{d-1}.$$

Their matrix $(H_{\boldsymbol{\alpha}})^T$ *corresponds to the Moore-Penrose pseudo-inverse of the matrix* $H_{\boldsymbol{\alpha}}$ *and is the analogue of our matrix* $(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T$ *in the proof of asymptotic normality of their estimator. We however show that the orthonormality assumption is not needed in the proofs.*

4.3.2 The efficient score estimator

In this section we extend the score approach of Section 4.3.1 by incorporating an estimate of the derivative of the link function ψ_0 to obtain an efficient estimator of $\boldsymbol{\alpha}_0$. Let $\hat{\psi}_{n,\boldsymbol{\alpha}}$ denote again the LSE of $\psi_{\boldsymbol{\alpha}}$ defined in Section 4.2 and define the estimate $\tilde{\psi}'_{nh,\boldsymbol{\alpha}}$ by

$$\tilde{\psi}'_{nh,\boldsymbol{\alpha}}(u) \stackrel{\text{def}}{=} \frac{1}{h}\int K\left(\frac{u-x}{h}\right)d\hat{\psi}_{n,\boldsymbol{\alpha}}(x),$$

where h is a chosen bandwidth. Here $d\hat{\psi}_{n,\boldsymbol{\alpha}}$ represents the jumps of the discrete function $\hat{\psi}_{n,\boldsymbol{\alpha}}$ and K is one of the usual symmetric twice differentiable kernels with compact support $[-1,1]$, used in density estimation. The estimator $\tilde{\boldsymbol{\alpha}}_n$ is given by

$$\tilde{\boldsymbol{\alpha}}_n \stackrel{\text{def}}{=} \mathbb{S}(\tilde{\boldsymbol{\beta}}_n), \tag{4.3.11}$$

where $\tilde{\boldsymbol{\beta}}_n$ is a zero crossing of $\xi_{2,nh}$ defined by

$$\xi_{2,nh}(\boldsymbol{\beta}) \stackrel{\text{def}}{=} \int \left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T \boldsymbol{x}\, \tilde{\psi}'_{nh,\boldsymbol{\alpha}}\big(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\big) \left\{y - \hat{\psi}_{n,\boldsymbol{\alpha}}\big(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\big)\right\} d\mathbb{P}_n(\boldsymbol{x}, y). \quad (4.3.12)$$

The function $\xi_{2,nh}$ is inspired by representing the sum of squares

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\right\}^2,$$

in a local coordinate system with $d-1$ unknown parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{d-1})^T$ followed by differentiation of the re-parametrized sum of squares w.r.t. $\boldsymbol{\beta}$ where we also consider differentiation of the function $\psi_{\boldsymbol{\alpha}}$.

We make the following additional assumptions for establishing the weak convergence of $\tilde{\boldsymbol{\beta}}_n$.

Assumptions A10-A11

A10. The function $\psi_{\boldsymbol{\alpha}}$ is two times continuously differentiable on $\mathcal{I}_{\boldsymbol{\alpha}}$ for all $\boldsymbol{\alpha}$.

A11. $(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T\, \mathbb{E}\Big[\psi_0'(\boldsymbol{\alpha}_0^T\boldsymbol{X})^2\ \text{Cov}(\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X})\Big]\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})$ is nonsingular.

Theorem 4.3.2. *Let Assumptions A1-A8, A10-A11 be satisfied. Let $\tilde{\boldsymbol{\alpha}}_n$ be defined by (4.3.11) and suppose $h \asymp n^{-1/7}$. Then*

(i) [Existence of a root] A crossing of zero $\tilde{\boldsymbol{\beta}}_n$ of $\xi_{2,nh}(\boldsymbol{\beta})$ exists with probability tending to one.

(ii) [Consistency]

$$\tilde{\boldsymbol{\alpha}}_n \xrightarrow{p} \boldsymbol{\alpha}_0, \qquad n \to \infty.$$

(iii) [Asymptotic normality] Define the matrices,

$$\tilde{\boldsymbol{A}} \stackrel{\text{def}}{=} \mathbb{E}\Big[\psi_0'(\boldsymbol{\alpha}_0^T\boldsymbol{X})^2\ Cov(\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X})\Big], \quad (4.3.13)$$

and

$$\tilde{\boldsymbol{\Sigma}} \stackrel{\text{def}}{=} \mathbb{E}\left[\left\{Y - \psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^2 \psi_0'(\boldsymbol{\alpha}_0^T\boldsymbol{X})^2 \left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T\right], \quad (4.3.14)$$

Then

$$\sqrt{n}(\tilde{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) \xrightarrow{d} N_d\left(\boldsymbol{0}, \tilde{\boldsymbol{A}}^-\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{A}}^-\right),$$

where $\tilde{\boldsymbol{A}}^-$ is the Moore-Penrose inverse of $\tilde{\boldsymbol{A}}$.

Remark 4.3.3. *The asymptotic variance of the estimator $\tilde{\boldsymbol{\alpha}}_n$ is similar to that obtained for the "efficient" estimates proposed in Xia and Härdle (2006) and in Kuchibhotla and Patra (2017). More details on the efficiency calculations can be found in Section 4.3.3, where we also illustrate that the asymptotic variance equals $\sigma^2\tilde{\boldsymbol{A}}^-$ in a homoscedastic model with var$(Y|\boldsymbol{X}=\boldsymbol{x})=\sigma^2$. This is the same as the inverse of $\mathbb{E}(\tilde{\ell}_{\boldsymbol{\alpha}_0,\psi_0}(\boldsymbol{X},Y)\tilde{\ell}_{\boldsymbol{\alpha}_0,\psi_0}(\boldsymbol{X},Y)^T)$ and, therefore, it follows that our estimator defined in (4.3.11), is efficient in the homoscedastic model. As also explained in Remark 2 of Kuchibhotla and Patra (2017), our estimator has moreover a high relative efficiency with respect to the optimal semiparametric efficiency bound if the constant variance assumption provides a good approximation of the truth.*

The asymptotic variance is obtained similarly to the derivations of the asymptotic limiting distribution for the simple score estimator as shown in Section 4.3.1. First the asymptotic variance is expressed in terms of the parametrization $\mathbb{S}$ as in (4.3.10) and next, similar to Lemma 4.3.1, equivalence to the expression $\tilde{\boldsymbol{A}}^-\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{A}}^-$ given in Theorem 4.3.2 is proved.

4.3.3 Efficient information in the single index model

The log likelihood of one observation in the SIM is given by

$$\ell_{\boldsymbol{\alpha},\psi}(\boldsymbol{x},y)=\log\left\{f_{\epsilon|\boldsymbol{X}}\left(y-\psi(\boldsymbol{\alpha}^T\boldsymbol{x})\right)g(\boldsymbol{x})\right\},$$

where $f_{\epsilon|\boldsymbol{X}}$ is the conditional density of ε given $\boldsymbol{X}=\boldsymbol{x}$ and g is the density of $\boldsymbol{X}$. The partial derivative w.r.t. $\boldsymbol{\alpha}$ of $\ell_{\boldsymbol{\alpha},\psi}$ is given by

$$\frac{\partial}{\partial\boldsymbol{\alpha}}\ell_{\boldsymbol{\alpha},\psi}(\boldsymbol{x},y)=\frac{\boldsymbol{x}\psi'(\boldsymbol{\alpha}^T\boldsymbol{x})f'_{\epsilon|\boldsymbol{X}}\left(y-\psi(\boldsymbol{\alpha}^T\boldsymbol{x})\right)}{f_{\epsilon|\boldsymbol{X}}\left(y-\psi(\boldsymbol{\alpha}^T\boldsymbol{x})\right)}.$$

Let $\{\psi_\eta:\eta\in(-1,1)\}$ be a path in the collection $\{\psi:\psi$ is increasing$\}$, differentiable w.r.t. η at $\eta=0$, and suppose

$$\psi_\eta=\psi \qquad \text{for } \eta=0,$$

and,

$$\frac{\partial}{\partial\eta}\psi_\eta(t)\Big|_{\eta=0}=a(t).$$

Then

$$\frac{\partial}{\partial\eta}\ell_{\boldsymbol{\alpha},\psi_\eta}(\boldsymbol{x},y)\Big|_{\eta=0}=\frac{a(\boldsymbol{\alpha}^T\boldsymbol{x})f'_{\epsilon|\boldsymbol{X}}\left(y-\psi(\boldsymbol{\alpha}^T\boldsymbol{x})\right)}{f_{\epsilon|\boldsymbol{X}}\left(y-\psi(\boldsymbol{\alpha}^T\boldsymbol{x})\right)}.$$

To obtain the efficient score function, we must solve the equation

$$\mathbb{E}\left[\left\{\frac{X_j\psi'(\boldsymbol{\alpha}^T\boldsymbol{X})f'_{\epsilon|\boldsymbol{X}}\left(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right)}{f_{\epsilon|\boldsymbol{X}}(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X}))} - \frac{a_{j,*}(\boldsymbol{\alpha}^T\boldsymbol{X})f'_{\epsilon|\boldsymbol{X}}\left(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right)}{f_{\epsilon|\boldsymbol{X}}(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X}))}\right\}\frac{a(\boldsymbol{\alpha}^T\boldsymbol{X})f'_{\epsilon|\boldsymbol{X}}\left(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right)}{f_{\epsilon|\boldsymbol{X}}(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X}))}\right]=0, \tag{4.3.15}$$

for an $\mathbb{R}^d$-valued function $a_* = (a_{1,*},\dots a_{d,*})^T$, where $a_*, a \in L_2^0(F)^d$ and for all $1 \le j \le d$. (see e.g. Huang (1996), p. 558, for similar computations with $\mathbb{R}$-valued function a_*). This amounts to solving in $a_{j,*}$:

$$\mathbb{E}\left[\frac{\left\{X_j\psi'(\boldsymbol{\alpha}^T\boldsymbol{X})-a_{j,*}(\boldsymbol{\alpha}^T\boldsymbol{X})\right\}f'_{\epsilon|\boldsymbol{X}}\left(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right)^2}{f_{\epsilon|\boldsymbol{X}}(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X}))^2}a(\boldsymbol{\alpha}^T\boldsymbol{X})\right]=0.$$

The efficient variance for $\boldsymbol{\alpha}$ in the single index model is derived in Newey and Stoker (1993), Delecroix et al. (2003) and Kuchibhotla and Patra (2017) among others. For the general case, we get that the efficient score function is given by

$$\tilde{\ell}_{\boldsymbol{\alpha},\psi}(\boldsymbol{x},y)=\frac{y-\psi\left(\boldsymbol{\alpha}^T\boldsymbol{x}\right)}{\sigma^2(\boldsymbol{x})}\psi'\left(\boldsymbol{\alpha}^T\boldsymbol{x}\right)\left\{\boldsymbol{x}-\frac{\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})\boldsymbol{X}|\boldsymbol{\alpha}^T\boldsymbol{X}=\boldsymbol{\alpha}^T\boldsymbol{x}\right\}}{\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X}=\boldsymbol{\alpha}^T\boldsymbol{x}\right\}}\right\}, \tag{4.3.16}$$

where $\sigma^2(\cdot)=\mathbb{E}(\varepsilon^2|\boldsymbol{X}=\cdot)$. We illustrate the derivation of this efficient score function in case that $\varepsilon|\boldsymbol{X}\sim N(0,\sigma^2(\boldsymbol{X}))$. We can write

$$\begin{aligned}
&\mathbb{E}\left[\frac{\left\{X_j\psi'(\boldsymbol{\alpha}^T\boldsymbol{X})-a_{j,*}(\boldsymbol{\alpha}^T\boldsymbol{X})\right\}f'_{\epsilon|\boldsymbol{X}}\left(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right)^2}{f_{\epsilon|\boldsymbol{X}}(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X}))^2}a(\boldsymbol{\alpha}^T\boldsymbol{X})\right]\\
&=\mathbb{E}\left[\left\{X_j\psi'(\boldsymbol{\alpha}^T\boldsymbol{X})-a_{j,*}(\boldsymbol{\alpha}^T\boldsymbol{X})\right\}\frac{\left\{y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right\}^2}{\sigma^4(\boldsymbol{X})}a(\boldsymbol{\alpha}^T\boldsymbol{X})\right]\\
&=\mathbb{E}\left[\mathbb{E}\left\{\frac{\left\{X_j\psi'(\boldsymbol{\alpha}^T\boldsymbol{X})-a_{j,*}(\boldsymbol{\alpha}^T\boldsymbol{X})\right\}}{\sigma^2(\boldsymbol{X})}\;\Big|\;\boldsymbol{\alpha}^T\boldsymbol{X}\right\}a(\boldsymbol{\alpha}^T\boldsymbol{X})\right].
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbb{E}\left\{\frac{\left\{X_j\psi'(\boldsymbol{\alpha}^T\boldsymbol{X})-a_{j,*}(\boldsymbol{\alpha}^T\boldsymbol{X})\right\}}{\sigma^2(\boldsymbol{X})}\;\Big|\;\boldsymbol{\alpha}^T\boldsymbol{X}\right\}\\
&=\psi'(\boldsymbol{\alpha}^T\boldsymbol{X})\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})X_j\;\Big|\;\boldsymbol{\alpha}^T\boldsymbol{X}\right\}-a_{j,*}(\boldsymbol{\alpha}^T\boldsymbol{X})\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})\;\Big|\;\boldsymbol{\alpha}^T\boldsymbol{X}\right\},
\end{aligned}$$

such that (4.3.15) is solved for,

$$a_*(u)=\psi'(u)\frac{\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})\boldsymbol{X}\;\Big|\;\boldsymbol{\alpha}^T\boldsymbol{X}=u\right\}}{\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})\;\Big|\;\boldsymbol{\alpha}^T\boldsymbol{X}=u\right\}}.$$

We conclude that the efficient score function for the semiparametric single index model if $\varepsilon|\boldsymbol{X} \sim N(0, \sigma^2(\boldsymbol{X}))$ is indeed given by

$$\tilde{\ell}_{\boldsymbol{\alpha},\psi}(\boldsymbol{x}, y) = \frac{y - \psi\left(\boldsymbol{\alpha}^T\boldsymbol{x}\right)}{\sigma^2(\boldsymbol{x})}\psi'\left(\boldsymbol{\alpha}^T\boldsymbol{x}\right)\left\{\boldsymbol{x} - \frac{\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})\boldsymbol{X}|\boldsymbol{\alpha}^T\boldsymbol{X} = \boldsymbol{\alpha}^T\boldsymbol{x}\right\}}{\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = \boldsymbol{\alpha}^T\boldsymbol{x}\right\}}\right\}.$$

4.4 Computation

In this section we describe how the score estimator $\hat{\boldsymbol{\alpha}}_n$ defined in (4.3.4) can be obtained using a local coordinate system representing the unit sphere in combination with a pattern search numerical optimization algorithm. An example of such a parameterization is the spherical coordinate system $\mathbb{S} : [0, \pi]^{(d-2)} \times [0, 2\pi] \mapsto \mathcal{S}_{d-1}$:

$$\begin{aligned}(\beta_1, \beta_2, \ldots, \beta_{d-1}) \mapsto &(\cos(\beta_1), \sin(\beta_1)\cos(\beta_2), \sin(\beta_1)\sin(\beta_2)\cos(\beta_3), \ldots,\\ &\sin(\beta_1)\ldots\sin(\beta_{d-2})\cos(\beta_{d-1}), \sin(\beta_1)\ldots\sin(\beta_{d-2})\sin(\beta_{d-1}))^T.\end{aligned}$$

The map parameterizing the positive half of the sphere $\mathbb{S} : \{(\beta_1, \beta_2, \ldots, \beta_{d-1}) \in [0, 1]^{(d-1)} : \|\boldsymbol{\beta}\| \leq 1\} \mapsto \mathcal{S}_{d-1}$:

$$(\beta_1, \beta_2, \ldots, \beta_{d-1}) \mapsto \left(\beta_1, \beta_2, \ldots, \beta_{d-1}, \sqrt{1 - \beta_1^2 - \ldots - \beta_{d-1}^2}\right)^T,$$

is another example that can be used provided α_d is positive. Prior knowledge about the position of $\boldsymbol{\alpha}_0$ can be derived from an initial estimate such as the LSE proposed in Balabdaoui et al. (2016).

We illustrate the set of equations corresponding to (4.3.3) for dimension $d = 3$ using the model

$$Y = \psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X}) + \varepsilon, \quad , \psi_0(\boldsymbol{x}) = x + x^3, \quad \alpha_{01} = \alpha_{02} = \alpha_{03} = 1/\sqrt{3},$$

$$X_1, X_2, X_3 \overset{i.i.d}{\sim} U[0, 1], \quad \varepsilon \sim N(0, 1),$$

where ε is independent of the covariate vector $\boldsymbol{X} = (X_1, X_2, X_3)^T$. For this model, we have

$$\boldsymbol{A} = (17/180)\boldsymbol{Q}, \quad \boldsymbol{\Sigma} = (1/36)\boldsymbol{Q} \quad \text{and} \quad \tilde{\boldsymbol{A}} = \tilde{\boldsymbol{\Sigma}} = 0.359656\boldsymbol{Q},$$

where

$$\boldsymbol{Q} = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix},$$

and where the matrices $\boldsymbol{A}, \boldsymbol{\Sigma}, \tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{\Sigma}}$ are defined in (4.3.7), (4.3.8), (4.3.13) and (4.3.14) respectively. Note that the rank of the matrices is equal to $d - 1 = 2$. We consider the parametrization

$$\mathcal{S}_3 = \{(\alpha_1, \alpha_2, \alpha_3) = (\cos(\beta_1)\sin(\beta_2), \sin(\beta_1)\sin(\beta_2), \cos(\beta_2)) : \\ 0 \le \beta_1 \le 2\pi, 0 \le \beta_2 \le \pi\}, \tag{4.4.1}$$

in $\mathbb{R}^2$ and we solve the problem

$$\begin{cases} s_1(\beta_1, \beta_2) = 0 \\ s_2(\beta_1, \beta_2) = 0 \end{cases} \tag{4.4.2}$$

where

$$s_1(\beta_1, \beta_2) = \frac{1}{n}\sum_{i=1}^{n}(-\sin(\beta_1)\sin(\beta_2)X_{i1} + \cos(\beta_1)\sin(\beta_2)X_{i2})\left\{Y_i - \hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\right\},$$

and

$$s_2(\beta_1, \beta_2) = \frac{1}{n}\sum_{i=1}^{n}(\cos(\beta_1)\cos(\beta_2)X_{i1} + \sin(\beta_1)\cos(\beta_2)X_{i2} - \sin(\beta_2)X_{i3}) \\ \cdot\left\{Y_i - \hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\right\}.$$

Note that

$$\mathbb{S}(\boldsymbol{\beta}_0) = (\cos(\beta_{01})\sin(\beta_{02}), \sin(\beta_{01})\sin(\beta_{02}), \cos(\beta_{02}))^T = \left(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}\right)^T,$$

and

$$\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0) = \begin{bmatrix} -\sin(\beta_{01})\sin(\beta_{02}) & \cos(\beta_{01})\cos(\beta_{02}) \\ \cos(\beta_{01})\sin(\beta_{02}) & \sin(\beta_{01})\cos(\beta_{02}) \\ 0 & -\sin(\beta_{02}) \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ 0 & -\sqrt{\frac{2}{3}} \end{bmatrix},$$

where $\beta_{01} = \pi/4$ and $\beta_{02} = \arctan(\sqrt{2})$. It can be easily seen from the above expression for the matrix $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)$ that the spherical coordinate system satisfies Assumption A8. We also have,

$$\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}) = (0, 0), \tag{4.4.3}$$

for all $\boldsymbol{\beta}$. This implies that the columns of $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})$ are perpendicular to the vector $\boldsymbol{\alpha} = \mathbb{S}(\boldsymbol{\beta})$. Note moreover that the columns are linearly independent and hence form a basis for $\{\boldsymbol{\alpha}\}^{\perp}$. Since the matrix

$$(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\,\mathbb{E}\Big[\psi_0'(\boldsymbol{\alpha}_0^T\boldsymbol{X})\,\mathsf{Cov}(\boldsymbol{X}|\boldsymbol{\alpha}_0^T\boldsymbol{X})\Big]\,(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)) = \begin{bmatrix} \frac{17}{90} & 0 \\ 0 & \frac{17}{60} \end{bmatrix},$$

Assumption A9 is satisfied. The asymptotic variance of $\hat{\boldsymbol{\alpha}}_n$ resp. $\tilde{\boldsymbol{\alpha}}_n$ defined in Theorem 4.3.1 resp. Theorem 4.3.2 is equal to,

$$\boldsymbol{A}^{-}\boldsymbol{\Sigma}\boldsymbol{A}^{-} = (100/289)\boldsymbol{Q} = 0.346021\boldsymbol{Q} \quad \text{and} \quad \tilde{\boldsymbol{A}}^{-}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{A}}^{-} = \tilde{\boldsymbol{A}}^{-} = 0.308937\boldsymbol{Q}.$$

None of the proposed criterion functions is convex and by the discontinuous nature of the functions s_1 and s_2, it is not possible to solve equations (4.4.2) exactly. This makes the computation of the estimators difficult. Standard optimization methods for convex loss functions cannot be used. The discreteness of the criterion functions moreover excludes methods that take derivative information into account since this derivative is often not defined. We search the crossing of zero (see Definition 2.4.1), by minimizing the sum of squares $s_1^2(\boldsymbol{\beta}) + s_2^2(\boldsymbol{\beta})$ over all possible values of $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$. Note that the crossing of zero of the score function is equivalent to the minimizer of the sum of squared component scores so that the minimization procedure is justified.
We use a derivative free optimization algorithm proposed by Hooke and Jeeves (1961). The method is a pattern-search optimization method that does not require the objective function to be continuous. The algorithm starts from an initial estimate of the minimum and looks for a better nearby point using a set of $2d$ equal step sizes along the coordinate axes in each direction, first making a step in the direction of the previous move. For the object function we take the sum of the squared values of the component functions, which achieves a minimum at a crossing of zero. If in no direction an improvement is found, the step size is halved, and a new search for improvement is done, with the reduced step sizes. This is repeated until the step size has reached a prespecified minimum. A very clear exposition of the method is given in Torczon (1997), section 4.3. In this paper also convergence proofs for the optimization algorithm are presented.

4.4.1 Lagrange approach

Instead of tackling the fact that our parameter space is essentially of dimension $d-1$ by the parametrization $\boldsymbol{\alpha} = \mathbb{S}(\boldsymbol{\beta})$ which locally maps $\mathbb{R}^{d-1}$ into the sphere $\mathcal{S}_{d-1}$, one can introduce the restriction $\|\boldsymbol{\alpha}\| = 1$ via a Lagrangian term. We then consider the problem of minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\right\}^2 + \lambda\left\{\|\boldsymbol{\alpha}\|^2 - 1\right\}, \tag{4.4.4}$$

where $\hat{\psi}_{n,\boldsymbol{\alpha}}$ is the LSE defined in Section 4.2 and λ is a Lagrange parameter which we add to the sum of squared errors to deal with the identifiability of the single-index model. We consider a Lagrange penalty for solving the optimization problem under the constraint that $\|\boldsymbol{\alpha}\| = 1$.

Analogously to the treatment given in Section 4.2, we next consider the derivative of (4.4.4) w.r.t. $\boldsymbol{\alpha}$, where we ignore the non-differentiability of the LSE $\hat\psi_{n,\boldsymbol{\alpha}}$. This leads to the set of equations,

$$\frac{1}{n}\sum_{i=1}^n X_{ij}\left\{\hat\psi_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i) - Y_i\right\} + \lambda\alpha_j = 0, \qquad 1 \le j \le d. \tag{4.4.5}$$

Here λ has to satisfy

$$\lambda = \lambda\sum_{j=1}^d \alpha_j^2 = -\frac{1}{n}\sum_{i=1}^n \boldsymbol{\alpha}^T\boldsymbol{X}_i\left\{Y_i - \hat\psi_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\right\}. \tag{4.4.6}$$

Plugging in the above expression for λ in (4.4.5), we would consider the score equation

$$\begin{aligned}\mathbf{0} &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\left\{Y_i - \hat\psi_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\right\} - \boldsymbol{\alpha}^T\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\left\{Y_i - \hat\psi_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_i)\right\}\right)\boldsymbol{\alpha}\\ &= \left(\boldsymbol{I} - \boldsymbol{\alpha}\boldsymbol{\alpha}^T\right)\int \boldsymbol{x}\left\{y - \hat\psi_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right\}\,d\mathbb{P}_n(\boldsymbol{x}, y),\end{aligned} \tag{4.4.7}$$

where $\boldsymbol{I}$ is the $d \times d$ identity matrix.

A computer program was implemented to solve (4.4.7). It has the advantage that we do not have to deal with the parametrization $\boldsymbol{\alpha} = \mathbb{S}(\boldsymbol{\beta})$, but has the disadvantage that we cannot assume that $\hat{\boldsymbol{\alpha}}_n$ has exactly norm 1 because we again have to deal with crossings of zero instead of exact equality to zero. One way to circumvent this problem is to normalize the solution after each iteration by dividing by its norm. This approach seems to provide reasonable solutions, although it is not entirely satisfactory from a theoretical point of view. Also note that if the right-hand side of (4.4.6) equals zero, so $\lambda = 0$, the equation does not force the norm of $\boldsymbol{\alpha}$ to be one; it only does so if $\lambda \neq 0$. Indeed, in our computer experiments, λ was never zero, so this problem did actually not occur, but λ will tend to zero with increasing sample sizes, so some numerical instability is to be expected.

For reasons of space we do not further describe all details of this approach, but instead restrict ourselves to showing a picture of the simple score estimate of ψ_0 for $n = 1,000$ and $d = 10$ for the simulation where all the X_i variables and the random error variable ε are standard normal and independent, $\psi_0(x) = x^3$ and $\boldsymbol{\alpha}_0 = (1/\sqrt{10}, \dots, 1/\sqrt{10})^T$. It is clear that the estimate of ψ_0 will be rather accurate because of the information provided by the 10 covariates X_i (instead of, say, just two covariates X_1, X_2).

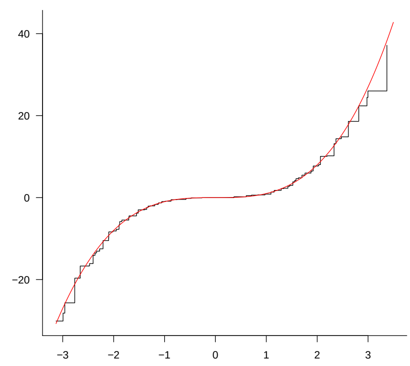


Figure 4.2: The real ψ_0 (red, solid) and the function $\hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}$ (black, step function) for $\psi_0(x) = x^3$, $\alpha_{0i} = 1/\sqrt{10}$, $X_i \overset{i.i.d}{\sim} N(0,1), i = 1 \dots 10$ and $n = 1,000$.

4.5 Simulations

In this section we illustrate the finite sample behavior of our single index score estimators proposed in Section 4.3.1 and Section 4.3.2. We also discuss in Section 4.5.2 the performance of other estimators for the SIM that avoid the use of smoothing techniques in the estimation approach and compare these estimators with our simple score estimator defined in Section 4.3.1. We use notations SSE and ESE to denote the simple respectively efficient score estimators of Section 4.3.1 respectively Section 4.3.2.

4.5.1 Finite sample behavior of the score estimators and the LSE

We consider the model

$$Y = \psi_0(\alpha_{01}X_1 + \alpha_{02}X_2) + \varepsilon = \left(\boldsymbol{\alpha}_0^T \boldsymbol{X}\right)^3 + \varepsilon, \tag{4.5.1}$$

where $\alpha_{0i} = 1/\sqrt{2}, i = 1,2$ and $\varepsilon \sim N(0,1)$, independent of $\boldsymbol{X}$. We consider two different distributions for the covariate vector $\boldsymbol{X}$, $X_i \overset{i.i.d}{\sim} U[1,2]$ and $X_i \overset{i.i.d}{\sim} N(0,1)$ for $i = 1,2$.

We consider the parametrization $\mathbb{S}(\beta_0) = (\cos(\beta_0), \sin(\beta_0))^T$ for these two-dimensional models. In each simulation setting, we estimate $\boldsymbol{\alpha}_0$ by the SSE and the ESE and compare the behavior of our proposed estimates with the LSE discussed in Section 4.2. For sample sizes $n = 100; 500; 1,000; 2,000; 5,000$ and $n = 10,000$ we generate $N = 5,000$ datasets from Model (4.5.1) and show, in Table 4.1 and Table 4.2, the mean and n times the covariance of the estimates. Table 4.1 and Table 4.2 also show the asymptotic values to which the results should converge based on Theorem 4.3.1 and Theorem 4.3.2. The limiting distribution of the LSE is still unknown and therefore no asymptotic results are provided for the LSE.

For all simulation studies, the results shown in Table 4.1 and Table 4.2 show convergence of n times the variance-covariance matrices towards its asymptotic values. The performance of the ESE is slightly better than the performance of the SSE; the difference between the asymptotic limiting variances is smaller in the model with Uniform$[1,2]$ covariates X_i than the difference in the model with standard normal covariates X_i. Although the model with standard normal covariates violates Assumptions A1, A2 and A4 given in Section 4.2, our proposed estimates perform reasonably well.

The behavior of the LSE is rather remarkable. Table 4.1 suggests an increase of n times the covariance matrix in contrast to Table 4.2 where n times the variance tends to stabilize. The results presented in Table 4.2 show that the performance of the LSE is clearly better than the performance of the SSE for small sample sizes when $X_i \sim N(0,1)$. For the model with uniform covariates, summarized in Table 4.1, our proposed score estimates outperform the LSE. The variances for the LSE presented in Table 4.1 and Table 4.2 suggest that the rate of convergence for the LSE is faster than the cube-root n-rate proved in Balabdaoui et al. (2016). The asymptotic distribution of the LSE needs to be addressed in further research.

4.5.2 Other estimators obtained without smoothing

In this section we compare the behavior of regression parameter estimators in the monotone single index model that are based on simple criterion functions and avoid the use of smoothing techniques. Although smoothing is necessary to obtain efficient estimators in the single index model, we want to point out that smoothing should not be the main concern when interest is in estimating the finite dimensional regression parameter. $\sqrt{n}$-consistent estimators with asymptotic normal limiting distribution with asymptotic variances that exceed the efficient variance have a good finite sample behavior. These simple estimators are often computationally more attractive than efficient estimators since efficiency is often based on smoothness conditions that are stronger than the conditions needed when smoothing techniques are avoided and consequently, efficient estimation algorithms require choosing one or several smoothing parameters.

The LSE, defined in Section 4.2 and the SSE, defined in Section 4.3.1, are both examples

Table 4.1: Simulation model ($X_i \sim U[1,2], d = 2$): The mean value ($\hat{\mu}_i =$ mean$(\hat{\alpha}_{in}), i = 1, 2$) and n times the variance-covariance ($\hat{\sigma}_{ij} = n{\cdot}cov(\hat{\alpha}_{in}, \hat{\alpha}_{jn})$,$i, j = 1, 2$) of the simple score estimate (SSE), the efficient score estimate (ESE) and the least squares estimate (LSE) for different sample sizes n with $N = 5,000$. The lines, preceded by ∞, give the asymptotic values.

Method	n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	$\hat{\sigma}_{12}$
SSE	100	0.707249	0.706389	0.040773	0.040866	-0.040787
	500	0.707224	0.706890	0.035018	0.035057	-0.035033
	1,000	0.707175	0.706992	0.033262	0.033273	-0.033265
	2,000	0.707173	0.707016	0.034017	0.034012	-0.034014
	5,000	0.707134	0.707070	0.034011	0.034012	-0.034011
	10,000	0.707109	0.707100	0.033344	0.033350	-0.033347
	∞	0.707107	0.707107	0.032439	0.032439	-0.032439
ESE	100	0.707293	0.706359	0.039631	0.039758	-0.039663
	500	0.707230	0.706888	0.033888	0.033922	-0.033900
	1,000	0.707185	0.706983	0.032302	0.032316	-0.032307
	2,000	0.707175	0.707015	0.032992	0.032989	-0.032990
	5,000	0.707130	0.707074	0.032925	0.032924	-0.032924
	10,000	0.707111	0.707098	0.032278	0.032283	-0.032280
	∞	0.707107	0.707107	0.031516	0.031516	-0.031516
LSE	100	0.706848	0.706624	0.052397	0.052415	-0.052321
	500	0.707060	0.707002	0.053547	0.053570	-0.053542
	1,000	0.707138	0.707000	0.053513	0.053573	-0.053535
	2,000	0.707122	0.707053	0.055502	0.055519	-0.055506
	5,000	0.707118	0.707079	0.059731	0.059756	-0.059741
	10,000	0.707128	0.707077	0.061843	0.061868	-0.061854
	∞	0.707107	0.707107	?	?	?

of estimators of the finite dimensional regression parameter $\boldsymbol{\alpha}_0$ that are based on the nonparametric LSE of the infinite dimensional nuisance parameter ψ_0. Inspired by the

Table 4.2: Simulation model ($X_i \sim N(0,1), d = 2$): The mean value ($\hat{\mu}_i = \text{mean}(\hat{\alpha}_{in}), i = 1,2$) and n times the variance-covariance ($\hat{\sigma}_{ij} = n{\cdot}\text{cov}(\hat{\alpha}_{in}, \hat{\alpha}_{jn}), i,j = 1,2$) of the simple score estimate (SSE), the efficient score estimate (ESE) and the least squares estimate (LSE) for different sample sizes n with $N = 5,000$. The lines, preceded by ∞, give the asymptotic values.

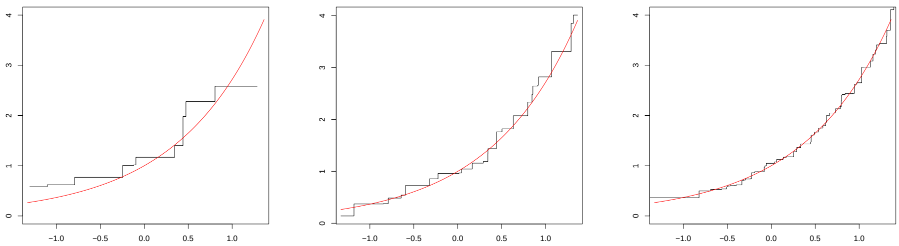
Method	n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	$\hat{\sigma}_{12}$
SSE	100	0.705202	0.706163	0.201378	0.200975	-0.200327
	500	0.706395	0.707509	0.109547	0.109291	-0.109371
	1,000	0.706824	0.707259	0.092557	0.092464	-0.092494
	2,000	0.707100	0.707056	0.080981	0.080966	-0.080967
	5,000	0.707026	0.707167	0.072021	0.071947	-0.071982
	10,000	0.707091	0.707113	0.067685	0.067665	-0.067674
	∞	0.707107	0.707107	0.055556	0.055556	-0.055556
ESE	100	0.706800	0.706173	0.087513	0.087878	-0.087480
	500	0.706905	0.707200	0.038450	0.038468	-0.038453
	1,000	0.706978	0.707190	0.031701	0.031672	-0.031685
	2,000	0.707061	0.707133	0.027930	0.027924	-0.027926
	5,000	0.707079	0.707128	0.023914	0.023907	-0.023911
	10,000	0.707104	0.707106	0.022827	0.022828	-0.022827
	∞	0.707107	0.707107	0.018519	0.018519	0.018519
LSE	100	0.706748	0.706135	0.093819	0.094319	-0.093715
	500	0.706710	0.707309	0.068561	0.068260	-0.068383
	1,000	0.706737	0.707389	0.061614	0.061325	-0.061459
	2,000	0.707009	0.707161	0.061123	0.061109	-0.061111
	5,000	0.707087	0.707110	0.060759	0.060722	-0.060738
	10,000	0.707074	0.707131	0.061708	0.061692	-0.061699
	∞	0.707107	0.707107	?	?	?

rank estimator proposed in Aragón and Quiroz (1995) for the current status model, we also propose a new estimator in this class and investigate its behavior via simulation studies.

To illustrate the criterion functions associated with the different estimators, we consider a simulated data sample from the model

$$Y = \exp(X_1/\sqrt{2} + X_2/\sqrt{2}) + \varepsilon, \qquad X_1, X_2 \sim U[-1,1] \text{ and } \varepsilon \sim N(0,1). \tag{4.5.2}$$

A picture of the LSE $\hat{\psi}_{n,\boldsymbol{\alpha}_0}$ for a sample of size $n = 100; 1,000$ and $n = 10,000$ is given in Figure 4.3.



(a) $n = 100$ **(b)** $n = 1,000$ **(c)** $n = 10,000$

Figure 4.3: The LSE $\hat{\psi}_{n,\boldsymbol{\alpha}_0}$ (black, step-wise) and the true function $\psi_0(x) = \exp(x)$ (red, solid) in model (4.5.2) for a sample of size (a) $n = 100$, (b) $n = 1,000$ and (c) $n = 10,000$.

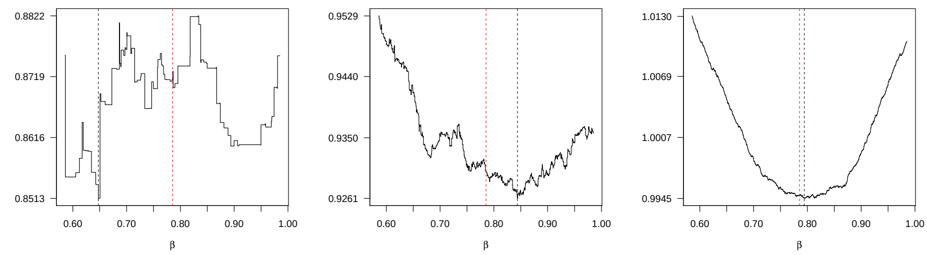
For the LSE and the SSE, figures of the criterion functions S_n defined in (4.2.1) and $\xi_{1,n}$, defined in (4.3.5), as a function of β, where β is defined by $(\alpha_1, \alpha_2) = (\cos(\beta), \sin(\beta))$ are given in Figure 4.4 and Figure 4.5 respectively. In model (4.5.2), the true parameter value $\beta_0 = \pi/4$.

4.5.2.1 The maximum rank correlation estimator (MRCE)

Han's maximum rank correlation estimator is motivated by the fact that $Y_i \geq Y_j$ is more likely than $Y_i < Y_j$ when $\boldsymbol{\alpha}_0^T \boldsymbol{X}_i \geq \boldsymbol{\alpha}_0^T \boldsymbol{X}_j$ if ψ_0 is increasing. The MRCE is defined by the maximizer of

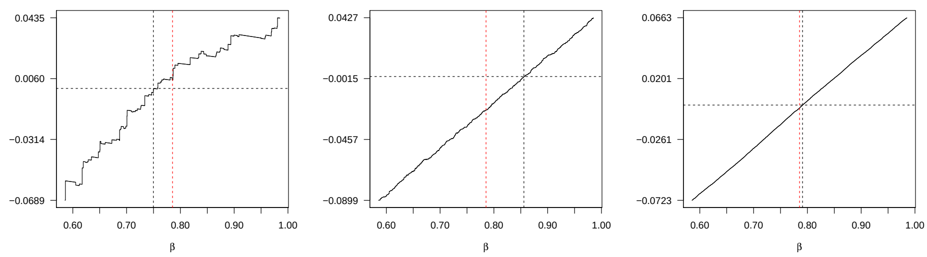
$$H_n(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \frac{1}{n(n-1)} \sum_{i \neq j} \{Y_i > Y_j\}\{\boldsymbol{\alpha}^T \boldsymbol{X}_i > \boldsymbol{\alpha}^T \boldsymbol{X}_j\}. \tag{4.5.3}$$

In contrast to the LSE and the SSE, estimation of the unknown link function ψ_0 is not considered with the MRCE.



(a) $n = 100$ **(b)** $n = 1,000$ **(c)** $n = 10,000$

Figure 4.4: The map $\beta \mapsto S_n((\cos(\beta), \sin(\beta))^T, \hat{\psi}_{n,\boldsymbol{\alpha}})$ (black, solid) in model (4.5.2) for a sample of size (a) $n = 100$, (b) $n = 1,000$ and (c) $n = 10,000$. The vertical reference lines indicate the position of the minimizer (black, dotted) and true $\beta_0 = \pi/4$ (red, dotted).



(a) $n = 100$ **(b)** $n = 1,000$ **(c)** $n = 10,000$

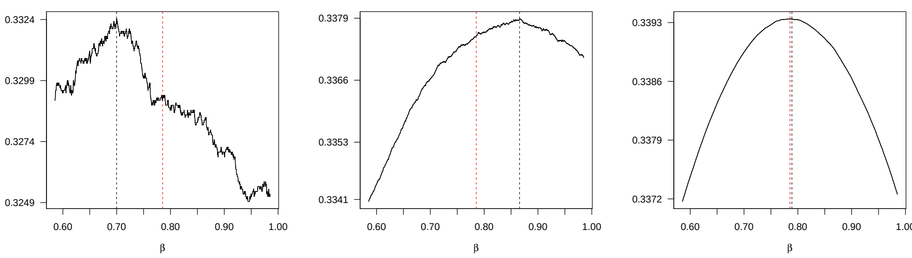
Figure 4.5: The map $\beta \mapsto \xi_{1,n}((\cos(\beta), \sin(\beta)^T))$ (black, solid) in model (4.5.2) for a sample of size (a) $n = 100$, (b) $n = 1,000$ and (c) $n = 10,000$. The vertical reference lines indicate the position of the zero-crossing (black, dotted) and true $\beta_0 = \pi/4$ (red, dotted).

4.5.2.2 The maximum rank estimator (MRE)

Inspired by the MRCE, Cavanagh and Sherman (1998) developed a new class of rank estimators defined by the maximizer of

$$R_n(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \frac{1}{n(n-1)} \sum_{i \neq j} M(Y_i)\{\boldsymbol{\alpha}^T \boldsymbol{X}_i > \boldsymbol{\alpha}^T \boldsymbol{X}_j\}, \tag{4.5.4}$$

where M denotes an increasing function on $\mathbb{R}$. In this section we investigate the behavior of the estimator when M is equal to the identity function, i.e. $M(y) = y$, and refer to this estimator as the maximum rank estimator. Since the responses in the binary choice model and the current status model are binary, it holds that the MRCE and the MRE are



(a) $n = 100$ **(b)** $n = 1,000$ **(c)** $n = 10,000$

Figure 4.6: The map $\beta \mapsto H_n((\cos(\beta), \sin(\beta)^T))$ (black, solid) in model (4.5.2) for a sample of size (a) $n = 100$, (b) $n = 1000$ and (c) $n = 10000$. The vertical reference lines indicate the position of the maximizer (black, dotted) and true $\beta_0 = \pi/4$ (red, dotted).

equivalent in these models. We illustrate this for the current status (cs) linear regression model and consider the MRE maximizing

$$R_{n,cs}(\boldsymbol{\alpha}) = \sum_{i \neq j} \Delta_i \{T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i > T_j - \boldsymbol{\alpha}^T \boldsymbol{X}_j\}.$$

This rank estimator is equivalent to Han's maximum rank correlation estimator, given by the maximizer of

$$\begin{aligned} H_{n,cs}(\boldsymbol{\alpha}) &= \sum_{i \neq j} \{\Delta_i > \Delta_j\}\{T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i > T_j - \boldsymbol{\alpha}^T \boldsymbol{X}_j\} \\ &= \sum_{i \neq j} \Delta_i (1 - \Delta_j)\{T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i > T_j - \boldsymbol{\alpha}^T \boldsymbol{X}_j\}. \end{aligned}$$

This can be seen as follows. Suppose that the observations are ordered in the $T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i$, i.e. $T_1 - \boldsymbol{\alpha}^T \boldsymbol{X}_1 \leq T_2 - \boldsymbol{\alpha}^T \boldsymbol{X}_2 \leq \ldots \leq T_n - \boldsymbol{\alpha}^T \boldsymbol{X}_n$. Then,

$$\begin{aligned} H_{n,cs}(\boldsymbol{\alpha}) &= \sum_{j<i} \Delta_i (1 - \Delta_j) = \sum_{j<i} \Delta_i - \sum_{j<i} \Delta_i \Delta_j = \sum_{j<i} \Delta_i - \frac{1}{2} \sum_{j \neq i} \Delta_i \Delta_j \\ &= R_{n,cs}(\boldsymbol{\alpha}) - \frac{1}{2} \sum_{j \neq i} \Delta_i \Delta_j. \end{aligned}$$

Since the second term in the expression above is independent of the ordering in $T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i$, the maximizers of $R_{n,cs}(\boldsymbol{\alpha})$ and $H_{n,cs}(\boldsymbol{\alpha})$ coincide and both estimators are equivalent.

The behavior of the map $\boldsymbol{\alpha} \mapsto H_n(\boldsymbol{\alpha})$ and the map $\boldsymbol{\alpha} \mapsto R_n(\boldsymbol{\alpha})$ are similar and we do not include pictures for the latter mapping.

4.5.2.3 The maximum rank estimator using the LSE of ψ_0 (LS-MRE)

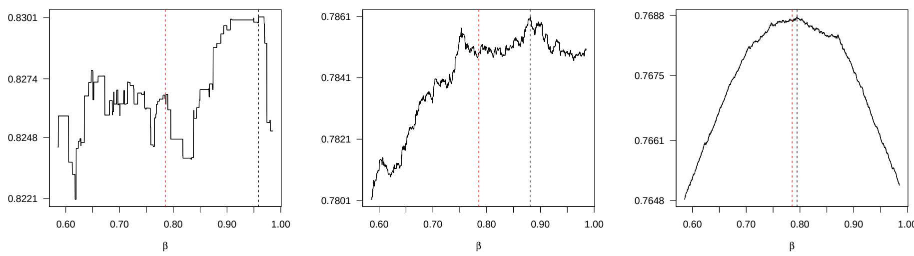
Aragón and Quiroz (1995) proposed two regression parameter estimators for the current status linear regression model based on the ranks of the observations $T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i$. The first estimator coincides with the MRE. The second estimator is defined by the maximizer of

$$\sum_{i \neq j} \hat{F}_{n,\boldsymbol{\alpha}}(T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i)\{T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i > T_j - \boldsymbol{\alpha}^T \boldsymbol{X}_j\},$$

where $\hat{F}_{n,\boldsymbol{\alpha}}$ is the MLE (see Chapter 2, Section 2.3) for fixed $\boldsymbol{\alpha}$. This motivates us to investigate the behavior of the regression parameter estimator for the monotone single index model, referred to as the LS-MRE, defined by the maximizer of

$$A_n(\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \frac{1}{n(n-1)} \sum_{i \neq j} \hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{X}_i)\{\boldsymbol{\alpha}^T \boldsymbol{X}_i > \boldsymbol{\alpha}^T \boldsymbol{X}_j\}, \tag{4.5.5}$$

where $\hat{\psi}_{n,\boldsymbol{\alpha}}$ is the LSE for fixed $\boldsymbol{\alpha}$. To the best of our knowledge this estimator has not been studied before and the asymptotic limiting distribution is still unknown. Since the LS-MRE is similar to the LSE, an M-estimator that involves the nonparametric LSE of ψ_0, it can be expected that similar theoretical issues appear when deriving the limiting behavior for both estimators.



(a) $n = 100$ **(b)** $n = 1,000$ **(c)** $n = 10,000$

Figure 4.7: The map $\beta \mapsto A_n((\cos(\beta), \sin(\beta)^T))$ (black, solid) in model (4.5.2) for a sample of size (a) $n = 100$, (b) $n = 1,000$ and (c) $n = 10,000$. The vertical reference lines indicate the position of the maximizer (black, dotted) and true $\beta_0 = \pi/4$ (red, dotted).

4.5.2.4 Asymptotic behavior

It has been shown in Section 4.3.1 for the SSE, in Sherman (1993) for the MRCE and in Cavanagh and Sherman (1998) for the MRE that these estimators are $\sqrt{n}$-consistent and have an asymptotic normal limiting distribution with asymptotic variance that is larger than the efficient variance. As pointed out in a footnote on p. 361 of Cavanagh and Sherman (1998) the expression for the asymptotic variance of the MRCE given in Theorem 4 of Sherman (1993) is only correct up to a factor 4. Unfortunately the same mistake for the MRE was made in the expression for the asymptotic variance of the MRE given in Theorem 2 of Cavanagh and Sherman (1998).

Although no proofs for the MRCE and the MRE have been published, we can prove that,

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{V}^{-}\boldsymbol{S}, \boldsymbol{V}^{-}) \qquad (4.5.6)$$

for specific choices of $\boldsymbol{V}$ and $\boldsymbol{S}$, where $\boldsymbol{V}^{-}$ is the Moore-Penrose inverse of $\boldsymbol{V}$. A sketch of the proof of (4.5.6) is given below. The reason that we again have to consider generalized inverses is that the normal limiting distributions are concentrated on the $(d-1)$-dimensional subspace, orthogonal to $\boldsymbol{\alpha}_0$ and are therefore degenerate. This is also clear from its covariance matrix $\boldsymbol{V}^{-}\boldsymbol{S}\boldsymbol{V}^{-}$, which is a matrix of rank $d-1$. The expressions for $\boldsymbol{V}$ and $\boldsymbol{S}$ are summarized in Table 4.3 for the monotone single index model and in Table 4.4 for the current status linear regression model. Here we introduce the notation g_0 to denote the density of the random variable $\boldsymbol{\alpha}_0^T\boldsymbol{X}$ resp. $T - \boldsymbol{\alpha}_0^T\boldsymbol{X}$.

Table 4.3: Asymptotic variances in the monotone single index model.

Method	$\boldsymbol{S}$
SSE	$\mathbb{E}\left[\left\{Y - \psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^2 \left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T\right]$
MRCE	$\mathbb{E}\left[\left\{2F_0(Y - \psi_0(\boldsymbol{\alpha}^T\boldsymbol{X})) - 1\right\}^2 \left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T g_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})^2\right]$
MRE	$\mathbb{E}\left[\left\{Y - \psi_0(\boldsymbol{\alpha}^T\boldsymbol{X})\right\}^2 \left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T g_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})^2\right]$
	$\boldsymbol{V}$
SSE	$\mathbb{E}\left[\psi_0'(\boldsymbol{\alpha}_0^T\boldsymbol{X}) \left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T\right]$
MRCE	$\mathbb{E}\left[2\psi_0'(\boldsymbol{\alpha}_0^T\boldsymbol{X})f_0(Y - \psi_0(\boldsymbol{\alpha}^T\boldsymbol{X}))\left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T g_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})\right],$
MRE	$\mathbb{E}\left[\psi_0'(\boldsymbol{\alpha}_0^T\boldsymbol{X})\left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T g_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})\right],$

Table 4.4: Asymptotic variances in the current status linear regression model.

Method	$\boldsymbol{S}$
SSE	$\mathbb{E}\left[F_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\{1-F_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\}\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}\mid T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}\mid T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T\right]$
MR(C)E	$\mathbb{E}\left[F_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\{1-F_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\}\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}\mid T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}\mid T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T g_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})^2\right]$
	$\boldsymbol{V}$
SSE	$\mathbb{E}\left[f_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}\mid\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}\mid\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T\right]$
MR(C)E	$\mathbb{E}\left[f_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}\mid\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}\mid\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T g_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\right]$

Asymptotic distribution of the MRCE and MRE:

Here we show that the asymptotic normal distribution for the MRCE is given by,

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)\xrightarrow{d}N_d(\mathbf{0},\boldsymbol{V}^-\boldsymbol{S},\boldsymbol{V}^-),$$

where $\boldsymbol{V}$ and $\boldsymbol{S}$ are defined in Table 4.3. A similar argument can be used to derive the asymptotic distribution of the MRE in terms of Moore-Penrose inverses. The asymptotic normality for the MRCE and the MRE are derived in Sherman (1993) and Cavanagh and Sherman (1998), where the authors restrict the parameter space to a compact subset $\{\boldsymbol{\alpha}\in\mathbb{R}^d:\alpha_d=1\}$. Each $\boldsymbol{\alpha}$ is represented as $(\beta,1)$ and only $d-1$ instead of d components are considered in the proofs of $\sqrt{n}$-consistency and asymptotic normality. Using the parametrization $\{\boldsymbol{\alpha}\in\mathbb{R}^d:\|\boldsymbol{\alpha}\|=1\}$ instead and considering a transformation $\mathbb{S}:B\subset\mathbb{R}^{d-1}\mapsto\{\boldsymbol{\alpha}\in\mathbb{R}^d:\|\boldsymbol{\alpha}\|=1\}$ as in Section 4.3, it follows, by similar arguments as in Sherman (1993) that for the MRCE we have

$$0\le H_n(\boldsymbol{\beta})-H_n(\boldsymbol{\beta}_0)=\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\mathbb{E}\left\{\nabla_2\tau((\boldsymbol{X},Y),\boldsymbol{\beta})\right\}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+\frac{1}{\sqrt{n}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{W}_n$$
$$+\,o_p(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2)+o_p(1/n),$$

where

$$\tau((\boldsymbol{x},y),\boldsymbol{\beta})=\mathbb{E}\left(\{y>Y\}\left\{\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}>\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}\right\}\right)+\mathbb{E}\left(\{Y>y\}\left\{\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}>\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right\}\right),$$

and where

$$\boldsymbol{W}_n=\frac{1}{\sqrt{n}}\sum_{i=1}^n\nabla_1\tau((\boldsymbol{X}_i,Y_i),\boldsymbol{\beta}_0),$$

converges in distribution to a Normal $N_{d-1}(\mathbf{0}, \boldsymbol{W})$ random vector. Here, ∇_i represents the ith partial derivative operator with respect to $\boldsymbol{\beta}$. Let $\boldsymbol{Q}$ denote $\mathbb{E}\left(\nabla_2\tau(\boldsymbol{X}, Y), \boldsymbol{\beta}\right)$. If $\boldsymbol{Q}$ is negative definite, than it follows by Theorem 2 in Sherman (1993) that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N_{d-1}(\mathbf{0}, \boldsymbol{Q}^{-1}\boldsymbol{W}\boldsymbol{Q}^{-1}).$$

Using an application of the delta-method, we conclude that,

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) \xrightarrow{d} N_d\left(\mathbf{0}, [\nabla_1\mathbb{S}(\boldsymbol{\beta}_0)][\boldsymbol{Q}^{-1}\boldsymbol{W}\boldsymbol{Q}^{-1}][\nabla_1\mathbb{S}(\boldsymbol{\beta}_0)]^T\right) = \left(\mathbf{0}, \boldsymbol{V}^{-}\boldsymbol{S}\boldsymbol{V}^{-}\right),$$

where $\boldsymbol{V}$ and $\boldsymbol{S}$ are given in Table 4.3 and where the last equality follows analogously to the proof of asymptotic normality of the SSE given in Appendix C.

The limiting distributions of the LSE and the LS-MRE are still unknown. Figure 4.4 and Figure 4.7 show a more irregular behavior of the criterion functions for the LSE and the LS-MRE compared to the smoother criterion functions for the SSE and the MRCE, shown in Figure 4.5 and Figure 4.6. Deriving the limiting distributions for the LSE and the LS-MRE is challenging. One of the difficulties arises from the non-differentiability of the LSE $\hat{\psi}_{n,\hat{\boldsymbol{\alpha}}}$ for ψ_0 appearing in the criterion functions S_n and A_n. This is, for example not the case with the efficient semiparametric LSE proposed in Ichimura (1993), where the criterion function is given by S_n defined in (4.2.1) but with $\hat{\psi}_{n\boldsymbol{\alpha}}$ replaced by a kernel estimate that is two times continuously differentiable with respect to $\boldsymbol{\alpha}$. By considering a Z-estimator instead of an M-estimator, this non-differentiability is somehow circumvented with the SSE. See also the discussion on the score approach given in Chapter 2, Section 2.2.

4.5.2.5 Finite sample behavior

To evaluate the finite sample behavior of the different estimators introduced in the previous Sections 4.5.2.1-4.5.2.3, we simulate $N = 5,000$ datasets from the model

$$Y = \psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X}) + \varepsilon, \tag{4.5.7}$$

where $\psi_0(x) = x + x^3$, $\alpha_{0i} = 1/\sqrt{3}, i = 1,2,3$ and $\varepsilon \sim N(0,1)$, independent of $\boldsymbol{X}$. We consider two different distributions for the covariate vector $\boldsymbol{X}$, $X_i \overset{i.i.d}{\sim} U[0,1]$ and $X_i \overset{i.i.d}{\sim} N(0,1)$ for $i = 1,2,3$.

Table 4.5 and Table 4.6 show the mean and n times the covariance matrix of the estimates for sample sizes $n = 100; 500; 1,000; 5,000$ and $n = 10,0000$ for the Uniform resp. Normal simulation setting. We calculated the asymptotic variances given in Table

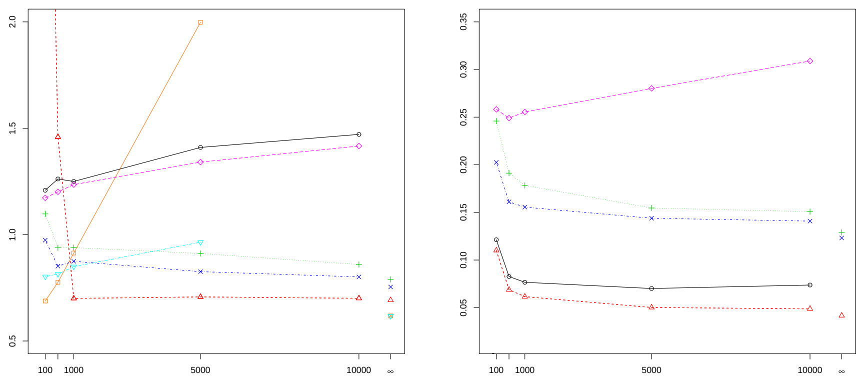
4.3 to which n times the covariance matrix should converge for the SSE, the MRCE and the MRE. We however note that only the Uniform model satisfies the assumptions needed to prove (4.5.6). The last column of Table 4.5 and Table 4.6 contains the distance between n times the covariance matrix of the estimates and the matrix $\boldsymbol{V}^{-}\boldsymbol{S}\boldsymbol{V}^{-}$ obtained by summing the squared distance of the corresponding matrix elements. The results for n times the variance of the estimates of α_3 are visualized in Figure 4.8.

For both simulation settings, the results show convergence of n times the covariance matrix towards the asymptotic values for the SSE, MRCE and MRE. The convergence rate is faster for the SSE than for the MRCE and MRE. We also note that the asymptotic values are smallest for the SSE in these models, with only a small difference for the Uniform setting but a larger difference in the Normal setting where the asymptotic values of the MRCE and MRE are substantially larger than the ones for the SSE.

For the LS-MRE, n times te covariance matrices increase with increasing sample size, suggesting a slower convergence rate than the parametric $\sqrt{n}$-rate for this estimator. Table 4.5 also shows a similar increase for the LSE in the Uniform model whereas a decrease of n times the covariance matrix for the LSE is shown in Table 4.6 for the Normal setting. The LSE even performs better than the MRCE and the MRE in the latter simulation model.

Finally, we also compared the inefficient estimates in the model with uniform covariates with the efficient penalized least squares estimate (PLSE) proposed by Kuchibhotla and Patra (2017) and the efficient EFM estimate proposed by Cui et al. (2011). The computation time of these efficient estimates is considerably longer than the time required for the score methods proposed in Section 4.3 and the estimates discussed in Section 4.5.2. Therefore, we do not report results for sample size $n = 10,000$ and simulated only N=2,500 data sets for the PLSE with $n = 5,000$. Boxplots of $\sum_{j=0}^{3}(\hat{\alpha}_j - \alpha_{0j})^2/3$, shown in Figure 4.9, illustrate that the PLSE and EFM estimate perform better than the SSE, MRCE and MRE for smaller sample sizes. As the sample size increases, the results for the efficient but computational intensive methods are no longer superior and the best performance is obtained with the SSE. The results for the PLSE and the EFM estimate depend furthermore on smoothing parameters which need to be selected carefully. Figure 4.8 clearly shows that n times the variance increases for the PLSE with increasing sample size, in contrast to the efficient convergence rate. This illustrates again that, in practice, methods that involve smoothing techniques are not necessarily a better choice than $\sqrt{n}$-consistent parameter-free methods, especially for larger sample sizes where the computation cost is enormous.

We conclude that it is worthwhile to consider parameter-free methods for estimation in the monotone single index model. The additional complexity (due to the smoothing parameter) does not necessarily result in better performances for efficient estimates. The increased computation time is only worthwhile when the sample size is small. The SSE is preferred for larger samples and moreover achieves better performances than the rank estimators (MRCE and MRE). The experiments in the normal model were in favor of the parametric $\sqrt{n}$-rate for the LSE whereas the uniform trials suggested a slower convergence rate. Even if the LSE leads at all to a $\sqrt{n}$-consistent estimate, its performance remains inferior to the score procedures in Section 4.3. Nevertheless, it remains an interesting topic to understand the behavior of the the LSE in the monotone single index model.



(a) **(b)**

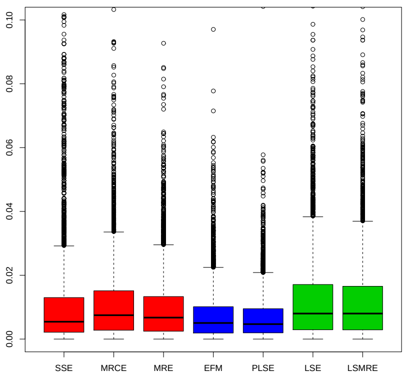
Figure 4.8: n times the variance of α_3 as a function of the sample size n for the simulation model with (a) $X_i \sim U[0,1]$ and (b) with $X_i \sim N(0,1)$ for the LSE (solid, black, $\circ$), SSE (dashed, red, $\triangle$), MRCE (dotted,green, $+$), MRE (dashed-dotted, blue, $\times$), LS-MRE (long-dashed, pink, $\diamond$), EFM (two-dashed, light blue, $\bigtriangledown$) and PLSE (solid, orange, $\square$). The points at ∞ represent the asymptotic values.

Table 4.5: Simulation model ($X_i \sim U[0,1], d = 3$): The mean value ($\hat{\mu}_i$ = mean($\hat{\alpha}_{in}$), $i = 1,2,3$), n times the variance-covariance ($\hat{\sigma}_{ij} = n\cdot$cov($\hat{\alpha}_{in}, \hat{\alpha}_{jn}$),$i,j = 1,2,3$) and the distance between n times the covariance matrix estimate and Σ for the least squares estimate (LSE), the simple score estimate (SSE), the maximum rank correlation estimate (MRCE), the maximum rank estimate (MRE) and the maximum rank estimate using the LSE of ψ_0 (LS-MRE) for different sample sizes n with $N = 5,000$. The lines, preceded by ∞, give the asymptotic values.

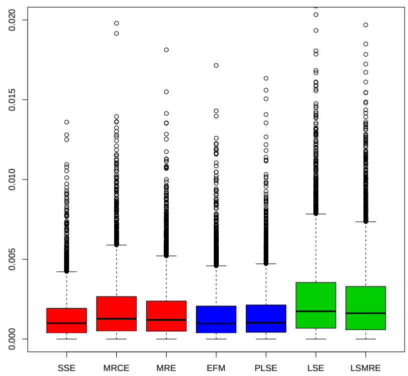
Method	n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	$\hat{\sigma}_{33}$	$\hat{\sigma}_{12}$	$\hat{\sigma}_{13}$	$\hat{\sigma}_{23}$	$d(\hat{\Sigma}, \Sigma)$
LSE	100	0.567232	0.566927	0.566318	1.201293	1.203355	1.208558	-0.577588	-0.592399	-0.577087	-
	500	0.575008	0.575510	0.574969	1.294027	1.227499	1.261569	-0.624327	-0.658712	-0.597269	-
	1,000	0.576605	0.575703	0.576451	1.276503	1.271816	1.249995	-0.645524	-0.628822	-0.618845	-
	5,000	0.577477	0.576993	0.576840	1.421290	1.446315	1.409902	-0.730242	-0.691148	-0.715894	-
	10,000	0.577146	0.577374	0.577146	1.506928	1.473131	1.471213	-0.755098	-0.751517	-0.718229	-
	∞	0.57735	0.57735	0.57735	?	?	?	?	?	?	
SSE	100	0.587614	0.541965	0.532872	1.544919	1.772945	4.386728	-0.955064	-1.496068	0.644483	25.896840
	500	0.573818	0.575772	0.576971	0.859601	0.844360	1.458978	-0.498872	-0.665879	-0.148044	4.603239
	1,000	0.574333	0.576839	0.579007	0.695695	0.753258	0.700203	-0.368076	-0.322984	-0.381043	2.976074
	5,000	0.576343	0.577253	0.578089	0.688215	0.709718	0.707268	-0.344059	-0.341827	-0.366363	2.914229
	10,000	0.576838	0.577328	0.577704	0.679286	0.708114	0.700672	-0.342635	-0.335454	-0.365785	2.891139
	∞	0.57735	0.57735	0.57735	0.692042	0.692042	0.692042	-0 .346021	-0.346021	-0.346021	
MRCE	100	0.567500	0.567568	0.568074	1.075928	1.137504	1.097447	-0.529870	-0.517154	-0.557786	5.485893
	500	0.576217	0.575649	0.575226	0.946660	0.974019	0.938209	-0.489812	-0.453586	-0.479446	4.613869
	1,000	0.576239	0.576586	0.576801	0.926655	0.931801	0.938936	-0.458336	-0.465244	-0.472174	4.499525
	5,000	0.577250	0.577172	0.577165	0.882133	0.881460	0.911412	-0.426668	-0.454751	-0.455531	4.270952
	10,000	0.577441	0.577291	0.577097	0.836557	0.856753	0.859607	-0.416975	-0.419253	-0.439996	4.048597
	∞	0.57735	0.57735	0.57735	0.789576	0.789576	0.789576	-0.394788	-0.394788	-0.394788	
MRE	100	0.568537	0.569242	0.568475	0.967113	1.016043	0.974079	-0.490072	-0.456418	-0.492633	4.567802
	500	0.576084	0.576064	0.575379	0.865385	0.890822	0.851881	-0.450847	-0.413482	-0.434273	3.977375
	1,000	0.576398	0.576643	0.576752	0.865576	0.864699	0.874510	-0.425514	-0.435668	-0.438601	3.978506
	5,000	0.577365	0.577191	0.577069	0.814612	0.820094	0.825936	-0.404390	-0.409542	-0.415915	3.729037
	10,000	0.577437	0.577292	0.577114	0.781254	0.813925	0.801210	-0.397353	-0.383830	-0.416803	3.623730
	∞	0.57735	0.57735	0.57735	0.753990	0.753990	0.753990	-0.376995	-0.376995	-0.376995	
LS-MRE	100	0.567857	0.567102	0.566043	1.158089	1.222715	1.172809	-0.579885	-0.539226	-0.600456	-
	500	0.575740	0.574947	0.575144	1.188863	1.194779	1.201250	-0.584415	-0.601986	-0.595143	-
	1,000	0.576197	0.575715	0.576921	1.241740	1.235253	1.235171	-0.618252	-0.614218	-0.617806	-
	5,000	0.577366	0.577222	0.576750	1.387179	1.379640	1.340894	-0.713591	-0.671955	-0.666950	-
	10,000	0.577212	0.577483	0.576991	1.394293	1.402226	1.416547	-0.689486	-0.704229	-0.712122	-
	∞	0.57735	0.57735	0.57735	?	?	?	?	?	?	

Table 4.6: Simulation model ($X_i \sim N(0,1), d = 3$): The mean value ($\hat{\mu}_i = \text{mean}(\hat{\alpha}_{in}), i = 1,2,3$), n times the variance-covariance ($\hat{\sigma}_{ij} = n{\cdot}\text{cov}(\hat{\alpha}_{in}, \hat{\alpha}_{jn}), i,j = 1,2,3$) and the distance between n times the covariance matrix estimate and Σ for the least squares estimate (LSE), the simple score estimate (SSE), the maximum rank correlation estimate (MRCE), the maximum rank estimate (MRE) and the maximum rank estimate using the LSE of ψ_0 (LS-MRE) for different sample sizes n with $N = 5,000$. The lines, preceded by ∞, give the asymptotic values.

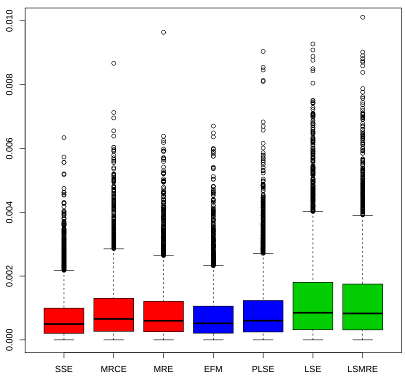
Method	n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	$\hat{\sigma}_{33}$	$\hat{\sigma}_{12}$	$\hat{\sigma}_{13}$	$\hat{\sigma}_{23}$	$d(\hat{\Sigma}, \Sigma)$
LSE	100	0.576253	0.575232	0.577442	0.117550	0.121344	0.121273	-0.057846	-0.059679	-0.061917	-
	500	0.576764	0.577339	0.577522	0.081738	0.081732	0.082715	-0.040379	-0.041181	-0.041484	-
	1,000	0.577343	0.577049	0.577455	0.081122	0.076944	0.076627	-0.040666	-0.040421	-0.036237	-
	5,000	0.577349	0.577424	0.577241	0.071714	0.071565	0.070109	-0.036610	-0.035116	-0.034965	-
	10,000	0.577243	0.577471	0.577318	0.072255	0.069788	0.073763	-0.034152	-0.038079	-0.035670	-
	∞	0.57735	0.57735	0.57735	?	?	?	?	?	?	
SSE	100	0.575519	0.575019	0.578488	0.113493	0.124740	0.110063	-0.062954	-0.051701	-0.058559	0.053908
	500	0.576568	0.577168	0.577949	0.068775	0.072997	0.068783	-0.036351	-0.032285	-0.036606	0.021234
	1,000	0.576909	0.577237	0.577743	0.062691	0.062724	0.061552	-0.031817	-0.030782	-0.030871	0.017500
	5,000	0.577193	0.577334	0.577497	0.049946	0.051876	0.050206	-0.025787	-0.024146	-0.026080	0.013040
	10,000	0.577218	0.577368	0.577452	0.047678	0.050129	0.048718	-0.024534	-0.023131	-0.025597	0.012451
	∞	0.57735	0.57735	0.57735	0.041667	0.041667	0.041667	-0.020833	-0.020833	-0.020833	
MRCE	100	0.574988	0.574617	0.575950	0.252922	0.249878	0.245900	-0.127500	-0.123730	-0.120269	0.256494
	500	0.577099	0.577073	0.576887	0.187896	0.193914	0.191217	-0.095492	-0.092522	-0.098216	0.165002
	1,000	0.577607	0.576964	0.577018	0.176862	0.177429	0.178333	-0.087992	-0.089192	-0.089015	0.147942
	5,000	0.577329	0.577313	0.577328	0.155366	0.157960	0.154613	-0.079387	-0.075988	-0.078577	0.124000
	10,000	0.577303	0.577355	0.577353	0.149451	0.153579	0.150804	-0.076088	-0.073316	-0.077506	0.119332
	∞	0.57735	0.57735	0.57735	0.128981	0.128981	0.128981	-0.064491	-0.064491	-0.064491	
MRE	100	0.575183	0.575212	0.576426	0.199617	0.200877	0.202508	-0.097920	-0.099979	-0.101862	0.177766
	500	0.577122	0.577191	0.576892	0.161150	0.166230	0.161093	-0.083228	-0.077936	-0.082881	0.131103
	1,000	0.577425	0.577094	0.577126	0.156489	0.156771	0.155526	-0.078799	-0.077816	-0.077693	0.124233
	5,000	0.577316	0.577328	0.577332	0.143959	0.145212	0.143865	-0.072627	-0.071307	-0.072569	0.112765
	10,000	0.577277	0.577379	0.577359	0.138858	0.144639	0.140865	-0.071304	-0.067543	-0.073327	0.110218
	∞	0.57735	0.57735	0.57735	0.123168	0.123168	0.123168	-0.061584	-0.061584	-0.061584	
LS-MRE	100	0.573239	0.575404	0.576779	0.249883	0.255367	0.258162	-0.120959	-0.125660	-0.132383	-
	500	0.576666	0.577033	0.577046	0.252064	0.252798	0.249036	-0.127917	-0.123899	-0.124663	-
	1,000	0.577145	0.577250	0.576976	0.268550	0.261884	0.255396	-0.137415	-0.131186	-0.124071	-
	5,000	0.577358	0.577305	0.577239	0.279751	0.298353	0.280200	-0.148836	-0.130687	-0.149572	-
	10,000	0.577360	0.577378	0.577235	0.295791	0.293497	0.308889	-0.140208	-0.155558	-0.153289	-
	∞	0.57735	0.57735	0.57735	?	?	?	?	?	?	



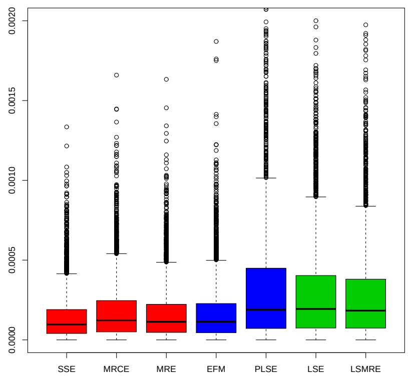
(a) $n = 100$



(b) $n = 500$



(c) $n = 1000$



(d) $n = 5000$

Figure 4.9: Boxplots of $\sum_{j=0}^{3}(\hat{\alpha}_j - \alpha_{0j})^2/3$ for the model with $X_i \sim U[0,1]$ for sample sizes (a) $n = 100$, (b) $n = 500$, (c) $n = 1000$ and (d) $n = 5000$. Red boxes correspond to $\sqrt{n}$-consistent but ineffcient methods (SSE, MRCE and MRE); blue boxes correspond to $\sqrt{n}-$consistent and effcient methods (EFM, PLSE) and green boxes correspond to methods with unknown limiting distribution (LSE, LS-MRE).

4.6 Real data example

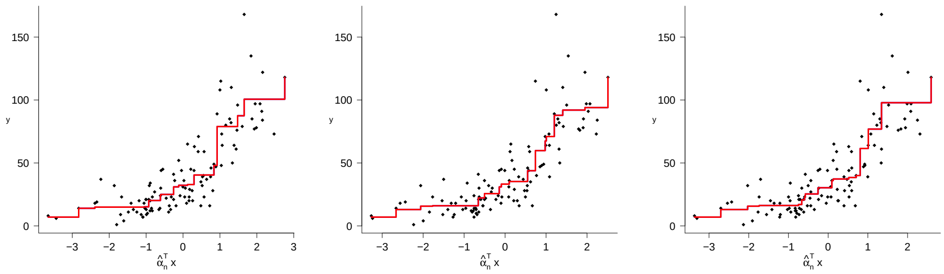
In this section we apply the estimation techniques on the Ozone data (Chambers et al., 1983). The data set contains observations on the ozone concentration for 153 consecutive days between May 1 and September 30, 1973. We study the relationship of the ozone concentration (Y) and the meteorological variables: solar radiation (R, Ly), temperature (T, $^\circ F$) and wind speed (W, mph) in a subset of the data consisting of 111 complete observations. For our data analysis we have scaled the covariates to have mean 0 and variance 1.

Table 4.7 summarizes the results of the regression parameter estimates for the LSE, SSE, MRCE, MRE and LS-MRE. The estimate $\hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}$ of ψ_0 together with a scatterplot

Table 4.7: Ozone data: Regression parameter estimates for the least squares estimate (LSE), the simple score estimate (SSE), the maximum rank correlation estimate (MRCE), the maximum rank estimate (MRE) and the maximum rank estimate using the LSE of ψ_0 (LS-MRE). $n = 111$. R = solar radiation, T = temperature and W = wind speed.

Method	R	T	W
LSE	0.261650	0.673180	-0.691641
SSE	0.288573	0.857762	-0.425406
MRCE	0.371694	0.833361	-0.409088
MRE	0.380572	0.835861	-0.395603
LS-MRE	0.269241	0.828638	-0.490783

of $(\hat{\boldsymbol{\alpha}}_n^T \boldsymbol{x}_i, y_i)$ is given in Figure 4.10 for the LSE, SSE and LS-MRE. We see that the estimates described in this paper result in similar estimated relationships between the ozone concentration and the meteorological variables.



(a) LSE **(b)** SSE **(c)** LS-MRE

Figure 4.10: Ozone data. Scatter plot $(\hat{\alpha}_n^T x_i, y_i)$ and $\hat{\psi}_{n,\hat{\alpha}_n}$ (red, step-function) for (a) the least squares estimate (LSE), (b) the simple score estimate (SSE) and (c) the maximum rank estimate using the LSE of ψ_0 (LS-MRE).

Part II

Varying coefficient models in a censored data setting using P-splines

Chapter 5

Penalized spline estimation in varying coefficient models with right censored data

Abstract

We propose a P-spline smoothing technique for the estimation of the varying coefficients in a varying coefficient model (VCM) with responses that are subject to random right censoring. Using the mean-preserving principle, we introduce two data-transformation approaches to transform the original censored observations into 'synthetic' observations, which are then used for the P-spline estimation. The synthetic response vector has the same expectation as the unobserved (due to censoring) response vector, conditional on covariates. Motivated by the research of Fan and Gijbels (1994), we first introduce model-independent transformations and later discuss, inspired by the approach of Buckley and James (1979), transformations that take the underlying regression model into account.

We give asymptotic support for the behavior of our proposed P-spline estimators and prove the consistency and asymptotic normality of our P-spline estimators for the coefficient functions in a VCM subject to random right censoring. Simulation studies compare its finite sample behavior with that of the smooth-backfitting estimator proposed by Yang et al. (2014) and illustrate good finite sample performance of our proposed P-spline estimates and moreover suggest improvements for the method proposed in Yang et al. (2014). Slightly better results are obtained with data transformations that take the true VCM into account. The latter transformation formulas require prior knowledge of the VCM which is obtained from the model-independent transformation methods. Based on simulations and real data examples, we conclude that the

combination of P-spline smoothing with a data transformation for censored observations is a good approach for estimating the coefficient functions in a VCM.

5.1 Introduction

Parametric regression models are commonly used for exploring relationships between a response variable and a set of explanatory variables. Linear models are often a good first approximation of the underlying association patterns but are sometimes not able to capture complex dynamic structures. An extension of the classical linear regression model is the varying coefficient model (VCM, Hastie and Tibshirani, 1993). These models are still linear in the regressors but with regression coefficients that are smooth functions in one or more other variables, considered as effect modifiers. VCMs have been used in a successful way in many applications, among which are longitudinal models (Hoover et al., 1998; Fan and Zhang, 2008), survival models (Cai et al., 2007; Ma and Wei, 2012), generalized regression models (Cai et al., 1999; Lee et al., 2012) and nonlinear time series (Cai et al., 2000). The most commonly used estimation methods for VCMs are kernel regression (Wu et al., 1998), polynomial splines (Huang et al., 2004) and smoothing splines (Hastie and Tibshirani, 1993). In this paper, we concentrate on the penalized spline (P-spline) smoothing technique proposed by Eilers and Marx (1996). P-spline regression is an extension of B-spline regression with a penalty in terms of finite differences of the coefficients of adjacent B-splines to protect against overfitting. P-splines are determined by the degree and the number and location of the knot points of the B-splines, the order of the difference penalty and a smoothing parameter. The consistency and asymptotic normality of the P-spline estimators for the regression coefficients in a VCM with longitudinal data was proven by Antoniadis et al. (2012).

Often encountered in the statistical analysis are situations where the response is not fully observed due to random right censoring, for example in medical and health care studies where patients leave the study for numerous reasons before the event of interest occurs (Lagakos et al., 1988, Nahman et al., 1992). Another example of censoring arises in reliability studies, where the failure time of a device might be censored if the device is still functional at the end of the experiment (Meeker, 1987). The popular proportional hazard model for right censored data (Cox, 1972) models the instantaneous risk as a product of a baseline hazard an an exponential factor. It models the relation between the response and covariates in an indirect way and is less simple to interpret than classical mean regression models, where interest is in direct modeling of the mean event time as a function of covariates. The accelerated failure time model (Wei, 1992) on the other hand does propose a direct linear relationship between the logarithm of the survival time and covariates, but unlike the Cox proportional hazard model, accelerated failure time models are often parametric and hence require additional assumptions on the underlying survival distribution. Ordinary least squares regression, which avoid specifying the distribution of the response variable for estimating the parameters in a linear regression model, needs however modification when some of the responses are not observed. Extensions of ordinary least squares to censored data settings were first considered by Buckley and James (1979). The estimation technique relies on constructing a synthetic response based on a transformation formula that is (conditional) mean preserving. The new response then replaces the original response in the ordinary least squares regression problem with complete data. The transformation studied by Buckley and James (1979) uses the underlying regression model and therefore needs an iterative estimation algorithm (see Jin et al. (2006) for the implementation of the iterative procedure). When transformed responses deal with transformations not depending on the unknown regression model but only on the censoring distribution, an iterative procedure is no longer needed at the cost however of increased variability in the transformed data. Transformations of this type were proposed by Koul et al. (1981), Zheng (1987), Fan and Gijbels (1994) and Leurgans (1987) among others. The combination of nonlinear mean regression models with synthetic data approaches for right censored data has mainly been studied for univariate covariates, see e.g. Fan and Gijbels (1994) and Heuchenne and Van Keilegom (2007). Recently more attention to multivariate regression models with right censored data transformation techniques is given by Yang et al. (2014) for the VCM and by Bravo (2014) for the varying coefficient partially linear model.

5.2 Model description

Consider the varying coefficient model

$$\begin{aligned} Y &= m(\mathbf{U},\mathbf{X}) + \sigma(\mathbf{U},\mathbf{X})\varepsilon \\ &= \beta_1(U_1)X_1 + \ldots + \beta_d(U_d)X_d + \sigma(U_1,X_1,\ldots,U_d,X_d)\varepsilon, \end{aligned} \tag{5.2.1}$$

where Y is the response variable, $\mathbf{U} = (U_1,\ldots,U_d)^T \in \mathcal{U}^d$ and $\mathbf{X} = (X_1,\ldots,X_d)^T \in \mathbb{R}^d$ are associated covariate vectors, where $\mathcal{U}^d$ denotes a d-dimensional interval on which the measurements are taken; ε is a mean-zero error term with variance one and (unknown) distribution function F, assumed to be independent of $\mathbf{U},\mathbf{X}$. The functions $\beta_1(u_1),\ldots,\beta_d(u_d)$ are the unknown regression coefficient functions at $\mathbf{U} = \mathbf{u} \equiv (u_1,\ldots,u_d)^T$ and $\sigma(\mathbf{u},\mathbf{x})$ is the variance of Y conditional on $\mathbf{U} = \mathbf{u}$ and $\mathbf{X} = \mathbf{x} \equiv (x_1,\ldots,x_d)^T$. When $X_1 \equiv 1$, the function β_1 is a nonzero intercept function representing the baseline effect.

We consider the case that the response Y of interest is subject to random right censoring. Let C be the censoring variable with survival function $G(\cdot|\mathbf{u},\mathbf{x})$ conditional on $(\mathbf{U},\mathbf{X}) = (\mathbf{u},\mathbf{x})$ and Δ be the censoring indicator $1_{\{Y \leq C\}}$. We observe a sample $(Z_i, \Delta_i, \mathbf{U}_i, \mathbf{X}_i), 1 \leq i \leq n$, from $(Z, \Delta, \mathbf{U}, \mathbf{X})$. We assume throughout that Y and C are independent, nonnegative continuous random variables.
In this paper we focus on estimating the regression curve $m(\mathbf{u},\mathbf{x})$. The estimation procedure for $\boldsymbol{\beta}(u) = (\beta_1(u_1), \ldots, \beta_d(u_d))^T$ consists of two steps: a mean-preserving data-transformation followed by P-spline smoothing using the transformed data. We describe the P-spline smoothing procedure with fully observed responses Y_i in Section 5.3 and describe in Section 5.4 two data-transformation approaches that allow a separation between the P-spline technique and the censored nature of the data.

5.3 P-spline estimator

Suppose that we have uncensored observations $(Y_i, \mathbf{U}_i, \mathbf{X}_i)$, $1 \leq i \leq n$. We use P-spline smoothing to estimate the varying coefficients in model (5.2.1). P-splines are an extension of regression splines with a penalty on the coefficients of adjacent B-splines. Each coefficient function β_p is approximated by a normalized B-spline basis expansion $\beta_p(u_p) \approx \sum_{l=1}^{m_p} B_{pl}(u_p; q_p)\alpha_{pl}$, where $\{B_{pl}(\cdot; q) : 1 \leq l \leq \ldots, K_p + q_p = m_p\}$ is the q_p-th degree B-spline basis using normalized B-splines such that $\sum_{l=1}^{m_p} B_{pl}(u_p; q_p) = 1$, with $K_p + 1$ equidistant knots $\boldsymbol{\xi}_p = (\xi_{p0}, \ldots, \xi_{pK_p})$. We use the notation $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_d^T)^T$ with $\boldsymbol{\alpha}_p = (\alpha_{p1}, \ldots, \alpha_{pm_p})^T$ for $1 \leq p \leq d$, to denote the unknown vector of B-spline regression coefficients and write $\mathcal{D} = \sum_{p=1}^{d} m_p$ for the dimension of $\boldsymbol{\alpha}$.
The P-spline optimization problem is given by

$$\begin{aligned}
&\min_{\boldsymbol{\alpha}} \left[\sum_{i=1}^{n} \left\{ Y_i - \sum_{p=1}^{d} X_{ip} \sum_{l=1}^{m_p} B_{pl}(U_i; q_p)\alpha_{pl} \right\}^2 + \sum_{p=1}^{d} \lambda_p \left(\sum_{l=k_p+1}^{m_p} (\Delta_p^k \alpha_{pl})^2 \right) \right] \\
&= \min_{\boldsymbol{\alpha}} \left\{ (\mathbf{Y} - \mathbf{R}\boldsymbol{\alpha})^T (\mathbf{Y} - \mathbf{R}\boldsymbol{\alpha}) + \alpha^T \mathbf{Q}_{\boldsymbol{\lambda}} \boldsymbol{\alpha} \right\}, \qquad (5.3.1)
\end{aligned}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{R} = (\mathbf{R}_1 | \ldots | \mathbf{R}_n)^T \in \mathbb{R}^{n \times \mathcal{D}}$ with $\mathbf{R}_i = (\mathbf{B}(\mathbf{U}_i))^T \mathbf{X}_i \in \mathbb{R}^{\mathcal{D} \times 1}$ and $\mathbf{B}(\mathbf{u}) \in \mathbb{R}^{d \times \mathcal{D}}$ given by

$$\mathbf{B}(\mathbf{u}) = \begin{pmatrix} B_{11}(u_1; q_0) & \ldots & B_{1m_1}(u_1; q_1) & 0 \ldots 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & \ddots & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 \ldots 0 & B_{d1}(u_d; q_d) & \ldots & B_{dm_d}(u_d; q_d) \end{pmatrix},$$

and

$$\mathbf{Q}_{\boldsymbol{\lambda}} = \operatorname{diag}(\lambda_1 \mathbf{D}_{k_1}^T \mathbf{D}_{k_1}, \ldots, \lambda_d \mathbf{D}_{k_d}^T \mathbf{D}_{k_d}) \in \mathbb{R}^{\mathcal{D} \times \mathcal{D}},$$

is a block diagonal matrix with $\lambda_p \mathbf{D}_{k_p}^T \mathbf{D}_{k_p}$ on the diagonal where $\mathbf{D}_{k_p}$ is the matrix representation of the k_p-th order difference operator Δ^{k_p}, i.e. $\Delta^{k_p}(\alpha_{pl}) = \sum_{h=0}^{k_p} (-1)^h \binom{k_p}{h} \alpha_{p(l-h)}$ (for $l \geq k_p$), with $k_p \in \mathbb{N}$; and $\lambda = (\lambda_1, \ldots, \lambda_d)$ is the vector of smoothing parameters satisfying $\lambda_p > 0, 1 \leq p \leq d$.

P-splines are computationally attractive since a closed form of the regression coefficient estimator exists. Antoniadis et al. (2012) showed that $\mathbf{R}^T\mathbf{R}+\mathbf{Q}_\lambda$ is invertible except on a set with probability tending to zero if $m_{\max}^{3/2}\lambda_{\max}/n = o(1)$, where $m_{\max} = \max(m_1, \ldots, m_d)$ and $\lambda_{\max} = \max(\lambda_1, \ldots, \lambda_d)$. Therefore the unique minimizer of $S(\boldsymbol{\alpha})$ is

$$\hat{\boldsymbol{\alpha}} \stackrel{\text{def}}{=} \left(\mathbf{R}^T\mathbf{R} + \mathbf{Q}_{\boldsymbol{\lambda}}\right)^{-1} \mathbf{R}^T \mathbf{Y}. \tag{5.3.2}$$

The P-spline estimator of $\boldsymbol{\beta}(\mathbf{u})$ is

$$\hat{\boldsymbol{\beta}}(\mathbf{u}) \stackrel{\text{def}}{=} \mathbf{B}(\mathbf{u})\hat{\boldsymbol{\alpha}} = (\hat{\beta}_1(u_1), \ldots, \hat{\beta}_d(u_d))^T, \text{ with } \hat{\beta}_p(u_p) = \sum_{l=1}^{m_p} B_{pl}(u_p; q_p)\hat{\alpha}_{pl}. \tag{5.3.3}$$

In Section 5.4, we construct, for randomly right censored data, a new response vector $\mathbf{Y}^*$ (the transformed response vector), that will replace $\mathbf{Y}$ in (5.3.2).

5.4 Data transformation approaches

We consider a data transformation approach and define the transformed response Y^* as

$$Y^* \stackrel{\text{def}}{=} \Delta\varphi(\mathbf{U}, \mathbf{X}, Z) + (1-\Delta)\psi(\mathbf{U}, \mathbf{X}, Z) = \begin{cases} \varphi(\mathbf{U}, \mathbf{X}, Z) & \text{if uncensored} \\ \psi(\mathbf{U}, \mathbf{X}, Z) & \text{if censored,} \end{cases}$$

with transformation functions φ and ψ so that

$$\mathbb{E}(Y^*|\mathbf{U}, \mathbf{X}) = \mathbb{E}(Y|\mathbf{U}, \mathbf{X}). \tag{5.4.1}$$

Condition (5.4.1) ensures that inference based on $(Y_i^*, \mathbf{U}_i, \mathbf{X}_i)$ preserves the conditional mean. In Section 5.4.1 we look at transformations that do not depend on the underlying regression model (5.2.1). Transformations that depend on model (5.2.1) are considered in Section 5.4.2. When a transformation depends on the unknown regression model, initial estimates for the regression curve and variance function are needed. In the second transformation method, we use as initial estimates for m and σ the estimates based on the model-independent transformation method of Section 5.4.1. We use the notation φ_1, ψ_1 and φ_2, ψ_2 to denote the transformation functions φ, ψ in methods one and two respectively.

5.4.1 Transformation method 1: model-independent transformations

From condition (5.4.1), we obtain the integral equation

$$\varphi_1(\mathbf{u},\mathbf{x},y)G(y|\mathbf{u},\mathbf{x}) - \int_0^y \psi_1(\mathbf{u},\mathbf{x},c)dG(c|\mathbf{u},\mathbf{x}) = y. \tag{5.4.2}$$

A specific class of solutions to (5.4.2), for all $y > 0, \mathbf{u} \in \mathcal{U}^d$ and $\mathbf{x} \in \mathbb{R}^d$ is given in Fan and Gijbels (1994), with $z > 0$ and $\gamma \in \mathbb{R}$,

$$\begin{aligned}\varphi_1(\mathbf{u},\mathbf{x},z) &= (1+\gamma)\int_0^z \frac{dt}{G(t|\mathbf{u},\mathbf{x})} - \gamma\frac{z}{G(z|\mathbf{u},\mathbf{x})},\\ \psi_1(\mathbf{u},\mathbf{x},z) &= (1+\gamma)\int_0^z \frac{dt}{G(t|\mathbf{u},\mathbf{x})}.\end{aligned} \tag{5.4.3}$$

The transformations only depend on the censoring distribution $G(\cdot|\mathbf{u},\mathbf{x})$ of C conditional on $(\mathbf{U},\mathbf{X}) = (\mathbf{u},\mathbf{x})$. Special cases of (5.4.3) are the methods proposed by Koul et al. (1981) and Leurgans (1987), taking $\gamma = -1$ and $\gamma = 0$ respectively. Since the functions φ_1 and ψ_1 depend on the unknown conditional survival function of C, an estimator $\hat{G}(\cdot|\mathbf{u},\mathbf{x})$ of $G(\cdot|\mathbf{u},\mathbf{x})$ is needed. A well-known problem with right censored data is, however, the estimation of a distribution function in the tail of the distribution. We therefore do not transform data points in the tail. As suggested in Fan and Gijbels (1994), we define

$$\begin{aligned}\hat{\varphi}_1(\mathbf{u},\mathbf{x},z) &\stackrel{\text{def}}{=} \bar{\varphi}_1(u,\mathbf{x},z)1_{\{z\le\tau_1(\mathbf{u},\mathbf{x})\}} + z1_{\{z>\tau_1(\mathbf{u},\mathbf{x})\}}\\ \hat{\psi}_1(\mathbf{u},\mathbf{x},z) &\stackrel{\text{def}}{=} \bar{\psi}_1(u,\mathbf{x},z)1_{\{z\le\tau_1(\mathbf{u},\mathbf{x})\}} + z1_{\{z>\tau_1(\mathbf{u},\mathbf{x})\}}\end{aligned}$$

for some $0 < \tau_1(\mathbf{u},\mathbf{x}) < \mathcal{T}(\mathbf{u},\mathbf{x}) = \sup\{t|H(z|\mathbf{u},\mathbf{x}) < 1\}$ with $H(z|\mathbf{u},\mathbf{x}) = P(Z \le z|\mathbf{U} = \mathbf{u}, \mathbf{X} = \mathbf{x})$ representing the distribution function of Z conditional on $(\mathbf{U},\mathbf{X}) = (\mathbf{u},\mathbf{x})$; where $\bar{\varphi}_1$ and $\bar{\psi}_1$ are given by (5.4.3) with G replaced by the estimator $\hat{G}$.
The synthetic response vector is defined as $\hat{\mathbf{Y}}_1^* = (\hat{Y}_{1i}^*,\ldots,\hat{Y}_{1n}^*)^T$ with, for $1 \le i \le n$,

$$\hat{Y}_{1i}^* \stackrel{\text{def}}{=} \Delta_i\hat{\varphi}_1(\mathbf{U}_i,\mathbf{X}_i,Z_i) + (1-\Delta_i)\hat{\psi}_1(\mathbf{U}_i,\mathbf{X}_i,Z_i),$$

and the P-spline estimator of $m(\mathbf{u},\mathbf{x})$ in method 1 is

$$\hat{m}_1(\mathbf{u},\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{x}^T\hat{\boldsymbol{\beta}}_1(\mathbf{u}) \quad \text{with} \quad \hat{\boldsymbol{\beta}}_1(\mathbf{u}) \stackrel{\text{def}}{=} \mathbf{B}(\mathbf{u})\big(\mathbf{R}^T\mathbf{R} + \mathbf{Q}_{\boldsymbol{\lambda}}\big)^{-1}\mathbf{R}^T\hat{\mathbf{Y}}_1^* \tag{5.4.4}$$

Remark 5.4.1. *In regression analysis, one is often interested in modeling $\mathbb{E}(f(Y)|\mathbf{U},\mathbf{X}) = m_f(\mathbf{U},\mathbf{X})$. For example, taking $f(y) = y$ gives model (5.2.1), and $f(y) = 1_{\{y\le t\}}$ corresponds to estimating the conditional distribution function of Y. It is possible to modify transformation functions φ_1 and ψ_1 such that we are estimating the* *conditional mean m_f, where f is a bounded nondecreasing function on $[0, \tau_1(\mathbf{u},\mathbf{x})]$, by defining the functions*

$$\hat{\varphi}_{1,f}(\mathbf{u},\mathbf{x},z) \stackrel{\text{def}}{=} \left\{(1+\gamma)\int_0^z \frac{df(t)}{\hat{G}(t|\mathbf{u},\mathbf{x})} - \gamma\frac{f(z)}{\hat{G}(z|\mathbf{u},\mathbf{x})}\right\} 1_{\{z\le\tau_1(\mathbf{u},\mathbf{x})\}} + f(z)1_{\{z>\tau_1(\mathbf{u},\mathbf{x})\}}$$

and,

$$\hat{\psi}_{1,f}(\mathbf{u},\mathbf{x},z) \stackrel{\text{def}}{=} \left\{(1+\gamma)\int_0^z \frac{df(t)}{\hat{G}(t|\mathbf{u},\mathbf{x})}\right\} 1_{\{z\le\tau_1(\mathbf{u},\mathbf{x})\}} + f(z)1_{\{z>\tau_1(\mathbf{u},\mathbf{x})\}}$$

and transformed responses

$$\hat{Y}^*_{1,f} \stackrel{\text{def}}{=} \Delta\hat{\varphi}_{1,f}(\mathbf{U},\mathbf{X},Z) + (1-\Delta)\hat{\psi}_{1,f}(\mathbf{U},\mathbf{X},Z). \tag{5.4.5}$$

The modified transformation formula is also suited for estimating the conditional variance of Y, i.e. $f(t) = (t - m(\mathbf{u},\mathbf{x}))^2$, when $\gamma = -1$, since for $\gamma = -1$, the nondecreasing condition for f is not necessary (see e.g. El Ghouch and Van Keilegom, 2008). As a consequence, when a (varying coefficient) model for $\sigma^2(\mathbf{U},\mathbf{X})$ is assumed, we can obtain a consistent estimate of $\sigma^2(\mathbf{u},\mathbf{x})$ by constructing

$$\hat{Y}^*_{1,\sigma^2} \stackrel{\text{def}}{=} \frac{\Delta(Z-\hat{m}_1(\mathbf{U},\mathbf{X}))^2}{\hat{G}(Z)} 1_{\{Z\le\tau_1(\mathbf{U},\mathbf{X})\}} + (Z-\hat{m}_1(\mathbf{U},\mathbf{X}))^2 1_{\{Z>\tau_1(\mathbf{U},\mathbf{X})\}}.$$

An estimate of $\sigma^2(\mathbf{u},\mathbf{x})$ is given by

$$\hat{\sigma}^2_1(\mathbf{u},\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{x}^T\mathbf{B}_{\sigma^2}(\mathbf{u})\big(\mathbf{R}^T_{\sigma^2}\mathbf{R}_{\sigma^2} + \mathbf{Q}_{\boldsymbol{\lambda},\sigma^2}\big)^{-1}\mathbf{R}^T_{\sigma^2}\hat{\mathbf{Y}}^*_{1,\sigma^2} \tag{5.4.6}$$

where the matrices $\mathbf{B}_{\sigma^2}, \mathbf{R}_{\sigma^2}$ and $\mathbf{Q}_{\boldsymbol{\lambda},\sigma^2}$ are the matrices $\mathbf{B}, \mathbf{R}$ and $\mathbf{Q}_{\boldsymbol{\lambda}}$ (introduced in Section 5.3) according to the model for σ^2. Another approach could be to estimate $\mathbb{E}(Y^2|\mathbf{U},\mathbf{X})$ and considering the difference $\mathbb{E}(Y^2|\mathbf{U},\mathbf{X}) - (\mathbb{E}(Y|\mathbf{U},\mathbf{X}))^2$. Note that we are not restricted to transformations with $\gamma = -1$ when we are estimating the conditional expectation of Y^2 since the function $f(t) = t^2$ in increasing on $\mathbb{R}^+$. Although the latter approach gives a consistent estimator of the variance function, in practice, numerical difficulties arise by taking the differences, since the difference is not guaranteed to be positive in finite samples.

5.4.2 Transformation method 2: model-dependent transformations

Based on the expression for the conditional expectation,

$$\mathbb{E}(Y|Z,\Delta,\mathbf{U},\mathbf{X}) = \Delta Z + (1-\Delta)\left\{m(\mathbf{U},\mathbf{X}) + \frac{\sigma(\mathbf{U},\mathbf{X})}{1-F\left(\frac{Z-m(\mathbf{U},\mathbf{X})}{\sigma(\mathbf{U},\mathbf{X})}\right)}\int_{(Z-m(\mathbf{U},\mathbf{X}))/\sigma(\mathbf{U},\mathbf{X})}^{\infty} tdF(t)\right\},$$

it follows that $\mathbb{E}\big(Y^*_{2[0]}|\mathbf{U},\mathbf{X}\big)=\mathbb{E}\big(Y|\mathbf{U},\mathbf{X}\big)$, for

$$Y^*_{2[0]} \overset{\text{def}}{=} \Delta\varphi^*_{2[0]}(\mathbf{U},\mathbf{X},Z)+(1-\Delta)\psi^*_{2[0]}(\mathbf{U},\mathbf{X},Z),$$

where $\varphi^*_{2[0]}(\mathbf{U},\mathbf{X},Z)\overset{\text{def}}{=} Z$ and

$$\psi^*_{2[0]}(\mathbf{U},\mathbf{X},Z)\overset{\text{def}}{=} m(\mathbf{U},\mathbf{X})+\frac{\sigma(\mathbf{U},\mathbf{X})}{1-F\left(\frac{Z-m(\mathbf{U},\mathbf{X})}{\sigma(\mathbf{U},\mathbf{X})}\right)}\int_{(Z-m(\mathbf{U},\mathbf{X}))/\sigma(\mathbf{U},\mathbf{X})}^{\infty} tdF(t).$$

In order to construct an estimator $\hat{\psi}_2$ of $\psi^*_{2[0]}$, we again consider a truncation device that avoids problems associated with the instability of an estimator for F. We follow the idea of Heuchenne and Van Keilegom (2007) and define ψ_2 and Y_2^* as follows:

$$\begin{aligned}\psi_2(\mathbf{U},\mathbf{X},Z) &\overset{\text{def}}{=} m(\mathbf{U},\mathbf{X})+\frac{\sigma(\mathbf{U},\mathbf{X})}{1-F(E^T)}\int_{E^T}^{S} tdF(t),\\ Y_2^* &\overset{\text{def}}{=} \Delta\varphi_2(\mathbf{U},\mathbf{X},Z)+(1-\Delta)\psi_2(\mathbf{U},\mathbf{X},Z),\end{aligned} \tag{5.4.7}$$

where the truncated residual $E^T=\min(E,S)$ with

$$E\overset{\text{def}}{=}\frac{Z-m(\mathbf{U},\mathbf{X})}{\sigma(\mathbf{U},\mathbf{X})} \quad \text{and} \quad S\overset{\text{def}}{=}\frac{\tau_2(\mathbf{U},\mathbf{X})-m(\mathbf{U},\mathbf{X})}{\sigma(\mathbf{U},\mathbf{X})},$$

for some $\tau_2(\mathbf{u},\mathbf{x})<\mathcal{T}(\mathbf{u},\mathbf{x})$. Let

$$\hat{E}\overset{\text{def}}{=}\frac{Z-\hat{m}_1(\mathbf{U},\mathbf{X})}{\hat{\sigma}_1(\mathbf{U},\mathbf{X})},\quad \hat{S}\overset{\text{def}}{=}\frac{\tau_2(\mathbf{U},\mathbf{X})-\hat{m}_1(\mathbf{U},\mathbf{X})}{\hat{\sigma}_1(\mathbf{U},\mathbf{X})} \quad \text{and} \quad \hat{E}^T\overset{\text{def}}{=}\min(\hat{E},\hat{S}).$$

We obtain the estimator $\hat{\psi}_2$ by replacing, in (5.4.7), m and σ by $\hat{m}_1$ and $\hat{\sigma}_1$, defined in (5.4.4) and (5.4.6), E^T and S by $\hat{E}^T$ and $\hat{S}$ and by replacing F by the Kaplan-Meier type estimator $\hat{F}$, constructed with residual observations $\hat{E}_i$, i.e.

$$\hat{F}(t)=1-\prod_{i:\hat{E}_i\le t}\left(1-\frac{1}{\sum_{j=1}^{n}1_{\{\hat{E}_j\ge\hat{E}_i\}}}\right)^{\Delta_i},$$

The transformed response vector $\hat{\mathbf{Y}}_2^*=(\hat{Y}^*_{21},\ldots,\hat{Y}^*_{2n})^T$ is defined by,

$$\hat{Y}^*_{2i}\overset{\text{def}}{=}\Delta_i Z_i+(1-\Delta_i)\hat{\psi}_2(\mathbf{U}_i,\mathbf{X}_i,Z_i). \tag{5.4.8}$$

The P-spline estimator $\hat{\boldsymbol{\beta}}_2(\mathbf{u})$ of $\boldsymbol{\beta}(\mathbf{u})$ in method 2 is obtained by replacing $\mathbf{Y}$ in (5.3.3) by $\hat{\mathbf{Y}}_2^*$.

Remark 5.4.2. *Note that, for method 1,* $\mathbb{E}(Y_1^*|\mathbf{U},\mathbf{X})= E(Y|\mathbf{U},\mathbf{X})$ *if* $Z\le\tau_1(\mathbf{U},\mathbf{X})$ *but for method 2 (as in Heuchenne and Van Keilegom (2007)),* $\mathbb{E}(Y_2^*|\mathbf{U},\mathbf{X})\neq$

$\mathbb{E}(Y|\mathbf{U},\mathbf{X})$, *since we truncate the integral in (5.4.7) and as a consequence we estimate a truncated mean* $\mathbb{E}(Y1_{\{\varepsilon\le S\}}|Z,\Delta,\mathbf{U},\mathbf{X})$. *The conditional expectation of* Y_2^* *will, however, be arbitrarily close to the conditional expectation of* Y *if* S *can be chosen arbitrarily close to* $\tau_F=\sup\{t|F(t)<1\}$, *which is possible when* $\tau_F\le\tau_J$, *where* J *is the distribution function of* $\{C-m(\mathbf{U},\mathbf{X})\}/\sigma(\mathbf{U},\mathbf{X})$ *and* $\tau_J=\sup\{t|J(t)<1\}$.

5.5 Asymptotic behavior

In Theorem 5.5.1, we show the consistency of the P-spline estimators obtained under transformation methods 1 and 2. The asymptotic normality of the estimators is considered in Theorem 5.5.2. Before stating the main results, we first give the following definition.

Definition 5.5.1. *Let* $\mathbb{G}(q_p,\boldsymbol{\xi}_p)$ *be the space of spline functions on* $\mathcal{U}_p$ *with fixed degree* q_p *and knot sequence* $\boldsymbol{\xi}_p$. *Let dist*$(\beta_p,\mathbb{G}(q_p,\boldsymbol{\xi}_p))=\inf_{g\in\mathbb{G}(q_p,\boldsymbol{\xi}_p)}\sup_{u\in\mathcal{U}}\mid\beta_p(u)-g(u)\mid$ *be the* L_∞*-distance between* β_p *and* $\mathbb{G}(q_p,\boldsymbol{\xi}_p)$. *The approximation error due to spline approximation is given by*

$$\rho_n\overset{\text{def}}{=}\max_{1\le p\le d}\textit{dist}(\beta_p,\mathbb{G}(q_p,\boldsymbol{\xi}_p)).$$

We use the notations $\hat{\boldsymbol{\beta}}_j=(\hat\beta_{j1},\ldots,\hat\beta_{jd})^T$, $\boldsymbol{\beta}_j^*=(\beta_{j1}^*,\ldots,\beta_{jd}^*)^T$ and $\tilde{\boldsymbol{\beta}}_j=(\tilde\beta_{j1},\ldots,\tilde\beta_{jd})^T$ for methods $j=1,2$, when we replace $\mathbf{Y}$ in expression (5.3.3) by $\hat{\mathbf{Y}}_j^*=(\hat Y_{j1}^*,\ldots,\hat Y_{jn}^*)^T$, $\mathbf{Y}_j^*=(Y_{j1}^*,\ldots,Y_{jn}^*)^T$, and $\mathbf{M}=(M_{j1},\ldots,M_{jn})^T$ with $M_{ji}=\mathbb{E}(Y_{ji}^*|\mathbf{U}_i,\mathbf{X}_i)$ for $1\le i\le n$ respectively. Note that $\mathbb{E}(\boldsymbol{\beta}_j^*|\mathcal{X}_n)=\tilde{\boldsymbol{\beta}}_j$ for $j=1,2$ where $\mathcal{X}_n=\{(\mathbf{U}_i^T,\mathbf{X}_i^T)^T,1\le i\le n\}$. See the Appendix for the definition of the L_2-distance and for Assumptions A-D in Theorems 5.5.1 and 5.5.2.

Theorem 5.5.1. *Suppose Assumptions A, B.1 and B.2 hold, then*

$$\|\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}\|_{L_2}=O_p\Bigg(n^{-1/2}m_{max}^{1/2}+n^{-1}m_{max}^{3/2}\lambda_{max}+\rho_n+\sup_{\mathbf{u},\mathbf{x}}\left\{\tau_1(\mathbf{u},\mathbf{x})\sup_{t\le\tau_1(\mathbf{u},\mathbf{x})}|\hat G(t|\mathbf{u},\mathbf{x})-G(t|\mathbf{u},\mathbf{x})|+\kappa(\mathbf{u},\mathbf{x})\right\}\Bigg).$$

where $\kappa(\mathbf{u},\mathbf{x})$ *is given by*

$$\max_{\phi=\varphi_1,\psi_1}\left[\mathbb{E}\left\{1_{\{Z>\tau_1(\mathbf{U},\mathbf{X})\}}|Z-\phi(\mathbf{U},\mathbf{X},Z)|\big|\mathbf{U}=\mathbf{u},\mathbf{X}=\mathbf{x}\right\}\right].$$

If, further Assumptions B.3 and C hold, then

$$\|\hat{\boldsymbol{\beta}}_2-\tilde{\boldsymbol{\beta}}_2\|_{L_2}=O_p\Bigg(n^{-1/2}m_{max}^{1/2}+n^{-1/2}\log n+n^{-1}m_{max}^{3/2}\lambda_{max}+\rho_n+ m_{max}^{-1/2}\left[\sup_{\mathbf{u},\mathbf{x}}\left\{\tau_1(\mathbf{u},\mathbf{x})\sup_{t\le\tau_1(\mathbf{u},\mathbf{x})}|\hat G(t|\mathbf{u},\mathbf{x})-G(t|\mathbf{u},\mathbf{x})|+\kappa(\mathbf{u},\mathbf{x})+\kappa_\sigma(\mathbf{u},\mathbf{x})\right\}\right]\Bigg).$$

where $\kappa_\sigma(\mathbf{u},\mathbf{x})$ is given by

$$\mathbb{E}\left\{1_{\{Z>\tau_1(\mathbf{U},\mathbf{X})\}}\left(Z-m(\mathbf{U},\mathbf{X},Z)\right)^2|1-\Delta/G(Z|\mathbf{U},\mathbf{X})|\big|\mathbf{U}=\mathbf{u},\mathbf{X}=\mathbf{x}\right\}.$$

Remark 5.5.1. *If $\sup_{\mathbf{u},\mathbf{x}}\kappa(\mathbf{u},\mathbf{x})\to 0$, the tail-contribution is negligible and the truncation device is justified. This condition was first introduced by Fan and Gijbels (1994) and suggests taking $\tau_1(\mathbf{u},\mathbf{x})$ as a sequence converging to $\mathcal{T}(\mathbf{u},\mathbf{x})$. If, e.g. conditional on $(\mathbf{U},\mathbf{X})=(\mathbf{u},\mathbf{x})$, Y ~Exp($\theta_{\mathbf{u},\mathbf{x}}$) and C ~Exp(ν) are independent exponentially distributed random variables, then $\kappa(\mathbf{u},\mathbf{x})=O\left(n^{-\theta_{u,x}}\log n\right)$ by taking $\tau_1(\mathbf{u},\mathbf{x})=\log n$ for all $\mathbf{u},\mathbf{x}$. As another illustration, suppose Y ~U$[0,\theta_{\mathbf{u},\mathbf{x}}]$ conditional on $(\mathbf{U},\mathbf{X})=(\mathbf{u},\mathbf{x})$ and C ~U$[0,\nu]$ are independent uniform random variables. After some tedious calculations we can show that $\kappa(\mathbf{u},\mathbf{x})\to 0$ for $\tau_1(\mathbf{u},\mathbf{x})=n^{-1}(n-1)\theta_{\mathbf{u},\mathbf{x}}$ and $\theta_{\mathbf{u},\mathbf{x}}\le\nu$. κ_σ arises similarly when method 1 is used to estimate σ using the transformation with $\gamma=-1$.*

Remark 5.5.2. *Suppose that each β_p is an r times continuously differentiable function $(1\le p\le d)$, if $q=q_p\ge r-1$, $m_{max}\asymp n^{1/(2r+1)}$ and $\lambda_{max}\asymp n^{\iota}$ with $\iota\le(r-1/2)/(2r+1)$, then $\|\beta_p^*-\beta_p\|_{L_2}=O_p(n^{-r/(2r+1)})$ reaches the optimal rate of convergence for nonparametric regression estimators. (Stone, 1992). The convergence rate of our P-spline estimator $\hat{\beta}_p^*$ is further influenced by the censored nature of the data.*

Theorem 5.5.2 gives the asymptotic normality results of the P-spline estimator. The variance-covariance matrix of $\boldsymbol{\beta}_j^*(\mathbf{u})$, conditional on $\mathcal{X}_n=\{(\mathbf{U}_i^T,\mathbf{X}_i^T)^T, 1\le i\le n\}$, is given by,

$$\mathbf{B}(\mathbf{u})\big(\mathbf{R}^T\mathbf{R}+\mathbf{Q}_{\boldsymbol{\lambda}}\big)^{-1}\Big(\sum_{i=1}^n\sigma_{j,i}^2\mathbf{R}_i\mathbf{R}_i^T\Big)\big(\mathbf{R}^T\mathbf{R}+\mathbf{Q}_{\boldsymbol{\lambda}}\big)^{-1}\mathbf{B}^T(\mathbf{u}), \tag{5.5.1}$$

where $\sigma_{j,i}^2=\mathrm{Var}(Y_{ji}^*|\mathbf{U}_i,\mathbf{X}_i)$.

Theorem 5.5.2. *If Assumptions A, B.1, B.2 and D.1 hold, then, for all $u_p\in\mathcal{U}_p$, $1\le p\le d$,*

$$\big(s.e.\left(\beta_{1,p}^*(u_p)\mid\mathcal{X}_n\right)\big)^{-1}\left(\hat{\beta}_{1,p}(u_p)-\beta_p(u_p)\right)\xrightarrow{d}N(0,1).$$

If Assumptions A, B, C and D.2 hold, then, for all $u_p\in\mathcal{U}_p$, $1\le p\le d$

$$\big(s.e.\left(\beta_{2,p}^*(u_p)\mid\mathcal{X}_n\right)\big)^{-1}\left(\hat{\beta}_{2,p}(u_p)-\tilde{\beta}_{2,p}(u_p)\right)\xrightarrow{d}N(0,1).$$

5.6 Practical aspects

5.6.1 Choice of the truncation points

We estimate the functional regression coefficients in VCM (5.2.1) by a combination of a data transformation for censored data and the P-spline estimator for complete case data.

The proposed data transformations involve an estimator of a distribution function. In the presence of censoring, nonparametric estimators of a distribution function are often inaccurate in the tail. To control this instability we use a truncation device that avoids the generation of transformed data in the tail.

In a clinical trial, censoring is often due to the termination of the study and hence not influenced by patient specific characteristics. In such situations the conditional survival function of C does not depend on the covariates, i.e. $G(\cdot|\mathbf{u},\mathbf{x}) \equiv G(\cdot)$, and the Kaplan-Meier product-limit estimator can be used to estimate the survival function of the censoring variable C. Note that, when estimating the censoring distribution G, the independent but not identically distributed event times $Y_i, 1 \leq i \leq n$ now play the role of censoring variables. For such situation the uniform strong consistency of the Kaplan-Meier estimator is still valid (see e.g. Zhou, 1991 and Bravo, 2014). If censoring is informative, but Y and C are conditionally independent given $\mathbf{U}, \mathbf{X}$, the conditional (on $\mathbf{U}, \mathbf{X}$) distribution of C, should be estimated in method 1, using, for example, the Beran (1981) estimator. However this may cause problems with the curse of dimensionality and one may want to consider a parametric or semi-parametric model for the censoring distribution instead.

In method 1, we do not transform data points when the observed response Z falls within the truncation area (τ_1, ∞). Choosing τ_1 too small implies that a lot of observations will not be transformed. On the other hand when τ_1 is chosen too large, large transformed responses are possible. In our numerical results we consider a censoring variable C independent of $(\mathbf{U}, \mathbf{X})$. We take $\tau_1 = \inf\{t|\hat{G}(t) < 0.01\}$ for method 1 and suggest to consider all jumps of the Kaplan-Meier estimator $\hat{F}$ in method 2 by taking $\hat{S} = \max(\hat{E}_1, \ldots, \hat{E}_n)$.

5.6.2 Smoothing parameter selection

Smoothing parameters are needed to control the amount of smoothing in the estimation process and imply a compromise between bias and variance. Undersmoothing arises by choosing too small values for the smoothing parameters, as a result, the bias will decrease at the price of an increased variance. When the smoothing parameters are too large, oversmoothing leads to a small variance but large bias (see Fahrmeir and Tutz, 2001, p. 187). Cross-validation (CV) is a popular parameter selection technique with complete case data based on minimizing the prediction error. With censored data, the prediction error cannot be calculated directly. We suggest to consider the transformed responses and choose the smoothing parameter $\boldsymbol{\lambda}$ that minimizes

$$CV(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \left\{ \frac{\hat{Y}_{ji}^{*} - \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_j(\mathbf{U}_i)}{1 - h_{ii}} \right\}^2,$$

where h_{ii} is the i-th diagonal element of the hat-matrix $H = \mathbf{R}(\mathbf{R}^T\mathbf{R} + \mathbf{Q}_{\boldsymbol{\lambda}})^{-1}\mathbf{R}^T$. The idea of using transformed responses in the prediction error calculation was also considered in Fan and Gijbels (1994) and Wang et al. (2008) among others. In practice, $CV(\boldsymbol{\lambda})$ is minimized over a d-dimensional grid of λ-values. With P-spline smoothing it is advisable to first consider a grid of the smoothing parameters on a logarithmic scale, which can later be fine-tuned when a more accurate smoothing parameter is desirable. Note that the P-spline estimator of β_p depends on the degree of the B-spline basis q_p, the number of knots $K_p + 1$, the order of the difference penalty k_p and the smoothing parameter λ_p. Cross-validation can be used to select several parameters, however, a good chosen smoothing parameter for fixed values of q_p, K_p and k_p will ensure a good fit. Cubic splines and a second order difference penalty are frequently used. A change in one of the parameters influences the choice of the other parameters, as a consequence, it is sufficient to select the smoothing parameters and keep the other parameters fixed.

5.6.3 Transformation parameter selection in method 1

The transformation parameter γ in method 1 determines the synthetic responses. We suggest to choose γ in a data-driven way. A cross-validation procedure can simultaneously select the smoothing parameter $\boldsymbol{\lambda}$ and transformation parameter γ when we search over a $(d+1)$-dimensional grid.

A second selection technique for the transformation parameter γ is based on the following observation. For $\gamma = -1$, all censored observations less than τ_1 are set equal to zero ($\psi_1 \equiv 0$), the uncensored observations are enlarged in order to compensate. If γ increases, we see that the variance of censored observations increases and that the enlargement of the uncensored observations is less pronounced (see Table 5.5). Therefore, we propose to select the transformation parameter γ that minimizes the sample variance of the transformed responses, denoted by the minimal-variance (MV) parameter γ_{MV}. Compared to CV-selection, the MV-selection procedure is appealing for being not computational intensive.

5.7 Finite sample behavior

In this section, we illustrate the finite sample behavior of our proposed P-spline estimates for VCMs when the observations are subject to random right censoring. Simulation studies are used to address the following objectives:

1. Compare our P-spline method with the smooth-backfitting (SBF) approach of Yang et al. (2014).

2. Investigate the quality of the data-transformation methods given in Section 5.4.

3. Evaluate the cross-validation selection criterion for the P-spline smoothing parameters.

We consider three different simulation scenarios. The first model is also used in Lee et al. (2012) and in Yang et al. (2014) and contrasts the performance between a spline smoothing and kernel approach for model-independent data transformation techniques. The second and third simulation model illustrate how model-dependent transformations increase the performance of model-independent approaches. The main difference between the two latter models is the nature of the random error terms which is homoscedastic in Model 2 and hetereoscedastic in Model 3. Therefore, Model 3 also gives insight in the quality of the variance estimation discussed in Remark 5.4.1. The simulation scenarios are as follows:

> **Model 1:** $Y = m(\mathbf{U},\mathbf{X}) + \sigma(\mathbf{U},\mathbf{X})\varepsilon = \beta_0(U_0) + \beta_1(U_1)X_1 + \beta_2(U_2)X_2 + \sigma(\mathbf{U},\mathbf{X})\varepsilon$, where $\beta_0(u) = 1 + \exp(2u-1)$, $\beta_1(u) = 0.5\cos(2\pi u), \beta_2(u) = u^2$ and $\sigma(\mathbf{U},\mathbf{X}) = 0.5 + (x_1^2+x_2^2)/(1+x_1^2+x_2^2)\exp(-2+(u_0+u_1)/2)$. The variables U_0, U_1, and U_2 are sampled from a Uniform$[0,1]$-distribution, the vector (X_1, X_2) is generated from a bivariate normal distribution with mean $(0,0)'$ and variance-covariance matrix $\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, and the random error has a normal distribution centered around 0 with standard deviation $\zeta = 1$ respectively $\zeta = 1.5$. The censoring variables are generated samples from a $\mathrm{N}(\mu_c, 1.5)$-distribution.

> **Model 2:** $Y = m(\mathbf{U},\mathbf{X}) + \varepsilon = \beta_1(U_1)X_1 + \beta_2(U_2)X_2 + \varepsilon$, where $\beta_1(u) = 2 + \sin(2\pi u)$, $\beta_2(u) = 1 + 0.1\exp(4u-1)$ with $U_1, U_2 \sim$ U$[0,1]$ and $(X_1, X_2)' \sim$ N$_2((3,3)'$, $\begin{pmatrix} 0.25 & -0.125 \\ -0.125 & 0.25 \end{pmatrix})$; ε has a standard normal distribution and the censoring variable has a uniform distribution on [6.5, R_c].

> **Model 3:** $Y = m(U,X) + \sigma(U)\varepsilon = \beta_0(U) + \beta_1(U)X + \sigma(U)\varepsilon$, where $\beta_0(u) = 2\exp(-2u-u^2))$, $\beta_1(u) = 1+5(u-0.5)^2$ and $\sigma^2(u) = \alpha\exp(-2u-0.4)/4$ where $\alpha = 1, 2$. We generate U from a Uniform$[0,1]$-distribution and X from a normal distribution with mean 1 and standard deviation 0.25; ε has a standard normal distribution and C is sampled from a $\mathrm{N}(\mu_c, 1)$-distribution

The parameters μ_c (in Models 1 and 3) and R_c (in Model 2) are chosen to control the level of censoring to $PC = 10\%, 30\%$ and 50%, respectively. No negative responses are observed in these simulation set-ups in correspondence to our model assumptions. We simulate 200 times a random sample of size $n = 250, 500$ from Models 2 and 3. For Model 1, we consider the exact same simulation settings as in Yang et al. (2014) and generate 500 samples of sizes $n = 200, 400$.

To evaluate the performance of the coefficient function estimates, we generate a uniform test sample $u_1, \ldots, u_{101}$ in $[0,1]$ for the random variables U_j and calculate the values for β_j and $\hat{\beta}_j$ in each simulation run. We then compute the relative error (RE) defined as (for $\hat{\beta}_j$),

$$\mathrm{RE}(\hat{\beta}_j) = \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_2 / \|\boldsymbol{\beta}_j\|_2,$$

with $\boldsymbol{\beta}_j = (\boldsymbol{\beta}_j(u_1), \ldots, \boldsymbol{\beta}_j(u_{101}))'; \hat{\boldsymbol{\beta}}_j = (\hat{\boldsymbol{\beta}}_j(u_1), \ldots, \hat{\boldsymbol{\beta}}_j(u_{101}))'$ and where $\| \cdot \|_2$ is the Euclidean distance. For the performance of the regression function estimate $\hat{m}$, we generate a test sample $\mathbf{x}_j = (x_{1j}, x_{2j})$, $1 \leq j \leq 101$, calculate $m_j = m(\mathbf{u}_j, \mathbf{x}_j)$ and $\hat{m}_j = \hat{m}(\mathbf{u}_j, \mathbf{x}_j)$, and compute the relative estimation error $\mathrm{RE}(\hat{m}) = \|\hat{\mathbf{m}} - \mathbf{m}\|_2 / \|\mathbf{m}\|_2$, where $\mathbf{m} = (m_1, \ldots, m_{101})$ and $\hat{\mathbf{m}} = (\hat{m}_1, \ldots, \hat{m}_{101})$. Tables 5.1- 5.2, Table 5.3 and Table 5.6 report the RE for the three simulation models introduced above.

We smooth each of the coefficient functions β_j with B-splines of degree 3 on 10 equidistant knots and use a penalty term with second order finite differences. The smoothing parameters λ_j are selected in a grid of size 8^d, where d equals the number of coefficient functions in the different simulation models. The CV-smoothing parameters (Section 5.6.2) are compared with optimal smoothing parameters that minimize the relative estimation error of the regression function m, referred to as the optimal selection criterion. Moreover, we present results for the smooth-backfitting estimates, where the optimal selection criterion is used to choose the bandwidths in a grid of equal size 8^d.

The simulation results, reported in Tables 5.1-5.6 and Figures 5.1-5.2, are discussed in the subsections below. The first objective is considered in Section 5.7.1. The importance of the transformation parameter selection in method 1 and the difference between model-dependent and model-independent transformations (objective 2) are outlined in Section 5.7.2. Section 5.7.3 addresses objective 3 and deals with the quality of the CV-smoothing parameter.

5.7.1 Comparison between P-spline and SBF-estimates

Yang et al. (2014) proposed a smoothing estimation approach for the VCMs with right censored responses. Their technique is a kernel analogue of the model-independent transformation method of Section 5.4.1 that combines an SBF-estimator with the transformation method proposed by Koul et al. (1981) using $\gamma = -1$. It is reasonable to compare our P-spline estimates using transformation method 1 with $\gamma = -1$ with the method proposed by Yang et al. (2014) since in both approaches the transformed response variable and covariates are the same. Table 5.1 and Table 5.2 therefore contrast the RE of a P-spline and kernel smoothing approach for the simulation scenario considered in Yang et al. (2014). The SBF-estimates of Yang et al. (2014) (SBF, M1$_K$)

perform often slightly better than the P-spline estimates with $\gamma = -1$ (P-SPLINE, $\text{M}1_K$) in Model 1. In Model 2, the P-spline estimates, however, outperform the SBF-estimates for PC = 10%, 30% (see Table 5.3).

In addition, we investigate the combination of an SBF-estimate with a data-driven MV-transformation parameter, instead of with the transformation proposed by Koul et al. (1981). The relative errors for both the P-spline and SBF-estimates decrease considerably if $\gamma = -1$ is changed to $\gamma = \gamma_{MV}$ (see Tables 5.1-5.3). We conclude from this decrease that the method proposed in Yang et al. (2014) can be improved if a different transformation parameter is considered. Interestingly, the choice between a P-spline smoothing or kernel smoothing approach has much less influence on the behavior of the estimates than the transformation parameter that is selected for the construction of the synthetic response. For the model-independent transformation methods, both combinations of a P-spline or SBF approach with a data-driven transformation parameter represent good choices for estimating the coefficient functions in the VCM under right censored observations. As expected, the relative errors in Tables 5.1-5.3 decrease with increasing sample size. On the contrary, an increase of the relative errors occurs if the percentages of censoring or the error variability increase.

From a theoretical point of view, both our P-spline and the SBF-estimates of Yang et al. (2014) converge at rate $n^{2/5}$ to a normal limiting distribution for suitably chosen smoothing parameters in case the censoring distribution is known and in case the coefficient functions are twice continuously differentiable (see Remark 5.5.2 and Lemma 1 in Yang et al., 2014). The difference between the true and estimated coefficient functions depends further on the approximation error of the censoring distribution for both P-spline and SBF-estimates. Hence, the choice between our P-spline approach and the SBF method of Yang et al. (2014) is hardly decided by the theoretical properties of the estimators. From a practical point of view, we note that P-spline estimates are obtained using simple matrix algebra whereas SBF-estimates require an iterative estimation procedure. The computations for the model-independent data transformation approaches took only a few seconds for the P-spline estimates and the computation time was, on average, 22 times larger for the SBF method than for the P-spline method in Simulation Model 2 (results not shown).

5.7.2 Findings on the transformation method

For the model-independent transformation method 1 of Section 5.4.1, Tables 5.1-5.3 show that a data-driven choice for the transformation parameter γ decreases the RE of the estimates compared to the choice $\gamma = -1$. Moreover, Table 5.3 shows how the estimates for transformation method 1 with the transformation by Koul et al. (1981) ($\gamma = -1$) behave worse than the estimates that are obtained when censoring is ignored (i.e. Z is considered as the true response). Consequently, we do not advise to use the transformation approach by Koul et al. (1981). Similar relative errors are obtained with the proposed data-driven transformations (MV and CV), with a slightly better result for the CV-method when the percentage of censoring is large. The computation cost for CV-selection is, however, considerably larger than for MV-selection. Therefore, we recommend to use the MV-transformation parameter when method 1 is used to obtain the synthetic response variable.

Table 5.3 and Table 5.6 report the performance of the model-dependent transformation method of Section 5.4.2 in case the initial starting estimates are obtained from the model-independent transformation method using $\gamma = \gamma_{MV}$. Transformation method 2 outperforms transformation method 1 for both the homoscedastic Model 2 and the heteroscedastic Model 3. Pointwise confidence bands of the P-spline estimates in Model 2 are illustrated in Figure 5.1. The curves show the 5% and 95% empirical quantiles at each grid point u_j and expose that the estimates obtained with method 2 are close to the true coefficient functions, even though in theory, method 2 is estimating a slightly different model. The results of method 2 are insensitive towards changes of γ in the initial transformation (results not shown). Additionally, Figure 5.1 shows once more the poor performance of the model-independent estimates using $\gamma = -1$.

5.7.3 Behavior of the smoothing parameter selection techniques

Table 5.4 presents the ratio of the relative error for m obtained with CV-selected smoothing parameters and optimal smoothing parameters in simulation Model 2 and illustrates that the CV-procedure works reasonably well (the ratio is close to one). Figure 5.2 presents scaled values of CV(λ_1, λ_2) and relative error of m for λ_1 and λ_2 (in Model 2) varying in $10^{\{0.5,0.6,\ldots,2.6\}}$ and demonstrates that the size of the CV-selected and optimal smoothing parameters are comparable. The behavior of both curves is similar. As a consequence the CV-method tends to select smoothing parameters that minimize the relative regression error for m. A data-driven bandwidth choice for the bandwidths of the SBF-estimates was proposed in Yang et al. (2014) and based on their results in Table 4 on p. 243, their comparison between the performance with optimal and data-driven bandwidth parameters is similar to our comparison in Table 5.4.

Table 5.1: Simulation Model 1: Average relative error for the estimates of the functions (F) $\beta_0, \beta_1, \beta_2$ and m obtained with the P-spline estimator and the smooth-backfitting estimator (SBF) with optimal smoothing parameters; using transformation method 1 (M1) (M1$_{MV}$: M1 with minimal-variability transformation, M$_K$: M1 with transformation by Koul et al. (1981) using $\gamma = -1$). n is the sample size, $\zeta = s.d.(\varepsilon)$ and PC is the percentage of censoring.

				P-SPLINE		SBF	
n	ζ	PC	F	M1$_{MV}$	M1$_K$	M1$_{MV}$	M1$_K$
200	1	10	β_0	0.0374	0.0728	0.0424	0.0721
			β_1	0.3303	0.7483	0.3555	0.7711
			β_2	0.2147	0.5440	0.2074	0.4799
			m	0.0742	0.1715	0.0781	0.1631
		30	β_0	0.0539	0.1517	0.0574	0.1334
			β_1	0.4443	1.3628	0.4856	1.3090
			β_2	0.3082	0.9922	0.2910	0.8652
			m	0.1039	0.3189	0.1072	0.2853
		50	β_0	0.0812	0.2594	0.0812	0.2232
			β_1	0.6312	2.0543	0.6832	1.8961
			β_2	0.4640	1.4777	0.4276	1.2512
			m	0.1543	0.4898	0.1530	0.4208
	1.5	10	β_0	0.0539	0.1020	0.0574	0.0938
			β_1	0.4490	0.9118	0.4802	0.9152
			β_2	0.3028	0.7186	0.2791	0.6345
			m	0.1034	0.2223	0.1049	0.2045
		30	β_0	0.0707	0.1903	0.0714	0.1658
			β_1	0.5513	1.5512	0.6005	1.4506
			β_2	0.3918	1.1901	0.3586	1.0379
			m	0.1323	0.3792	0.1319	0.3330
		50	β_0	0.1010	0.3069	0.0938	0.2648
			β_1	0.7315	2.2477	0.7752	2.0514
			β_2	0.5446	1.7031	0.4926	1.4363
			m	0.1836	0.5545	0.1746	0.4740

Table 5.2: Simulation Model 1: Average relative error for the estimates of the functions (F) $\beta_0, \beta_1, \beta_2$ and m obtained with the P-spline estimator and the smooth-backfitting estimator (SBF) with optimal smoothing parameters; using transformation method 1 (M1) (M1$_{MV}$: M1 with minimal-variability transformation, M$_K$: M1 with transformation by Koul et al. (1981) using $\gamma = -1$). n is the sample size, $\zeta = s.d.(\varepsilon)$ and PC is the percentage of censoring.

				P-SPLINE		SBF	
n	ζ	PC	F	M1$_{MV}$	M1$_K$	M1$_{MV}$	M1$_K$
400	1	10	β_0	0.0265	0.0490	0.0316	0.0529
			β_1	0.2433	0.5631	0.2693	0.5855
			β_2	0.1581	0.3923	0.1559	0.3491
			m	0.0539	0.1249	0.0583	0.1217
		30	β_0	0.0374	0.1127	0.0424	0.0990
			β_1	0.3268	1.0173	0.3599	1.0072
			β_2	0.2238	0.7260	0.2159	0.6332
			m	0.0755	0.2385	0.0794	0.2152
		50	β_0	0.0592	0.1954	0.0608	0.1706
			β_1	0.4839	1.4747	0.5142	1.3873
			β_2	0.3342	1.2063	0.3127	1.0467
			m	0.1145	0.3803	0.1145	0.3332
	1.5	10	β_0	0.0387	0.0700	0.0436	0.0693
			β_1	0.3360	0.6946	0.3606	0.7205
			β_2	0.2234	0.5107	0.2066	0.4508
			m	0.0762	0.1612	0.0787	0.1530
		30	β_0	0.0500	0.1459	0.0539	0.1292
			β_1	0.4177	1.2250	0.4506	1.1785
			β_2	0.2851	0.9122	0.2627	0.7905
			m	0.0975	0.2958	0.0985	0.2627
		50	β_0	0.0735	0.2421	0.0721	0.2059
			β_1	0.5633	1.7462	0.5967	1.5620
			β_2	0.3960	1.4647	0.3561	1.2511
			m	0.1364	0.4618	0.1319	0.3909

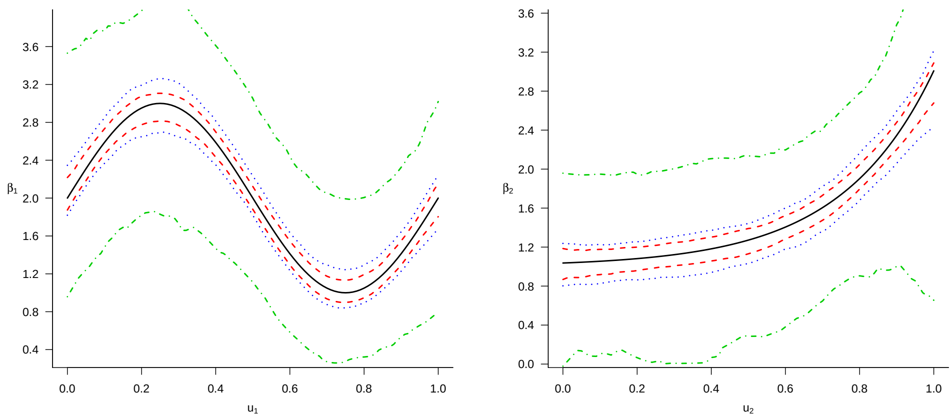
Table 5.3: Simulation Model 2: Average relative error for the estimates of the functions (F) β_1, β_2 and m obtained with the P-spline estimator and the smooth-backfitting estimator (SBF) with optimal smoothing parameters; using transformation methods 1 (M1) and 2 (M2). M1$_{CV}$: M1 with cross-validation transformation, M1$_{MV}$: M1 with minimal-variability transformation, M$_K$: M1 with transformation by Koul et al. (1981) using $\gamma = -1$. M$_Z$ indicates the estimator when no transformation is applied to the observed response (Z, Δ). n is the sample size, PC is the percentage of censoring.

			P-SPLINE					SBF	
n	PC	F	M1$_{CV}$	M1$_{MV}$	M1$_K$	M2	M$_Z$	M1$_{MV}$	M1$_K$
250	10	β_1	0.0514	0.0519	0.1718	0.0424	0.0778	0.0806	0.2127
		β_2	0.0674	0.0680	0.2218	0.0564	0.0876	0.0983	0.2659
		m	0.0260	0.0262	0.0740	0.0229	0.0474	0.0355	0.0840
	30	β_1	0.0856	0.0868	0.3522	0.0546	0.1704	0.1164	0.4003
		β_2	0.1121	0.1131	0.4621	0.0730	0.1580	0.1430	0.4853
		m	0.0414	0.0419	0.1637	0.0303	0.1135	0.0521	0.1675
	50	β_1	0.1245	0.1322	0.7222	0.0862	0.2741	0.1828	0.7120
		β_2	0.1612	0.1716	0.9379	0.1137	0.2266	0.2029	0.8216
		m	0.0684	0.0742	0.3867	0.0571	0.1820	0.0830	0.3409
500	10	β_1	0.0367	0.0366	0.1230	0.0301	0.0641	0.0653	0.1549
		β_2	0.0482	0.0481	0.1604	0.0394	0.0724	0.0783	0.1961
		m	0.0188	0.0190	0.0557	0.0157	0.0433	0.0274	0.0679
	30	β_1	0.0608	0.0605	0.2713	0.0361	0.1612	0.0914	0.3172
		β_2	0.0800	0.0796	0.3598	0.0484	0.1493	0.1092	0.3925
		m	0.0301	0.0300	0.1276	0.0207	0.1106	0.0411	0.1412
	50	β_1	0.0997	0.1070	0.6291	0.0667	0.2682	0.1552	0.6224
		β_2	0.1365	0.1460	0.7931	0.0923	0.2240	0.1722	0.6943
		m	0.0539	0.0620	0.3652	0.0465	0.1812	0.0716	0.3146

average relative error based on true (unobserved) responses with the P-spline estimate for: $n = 250$: β_1: 0.0400; β_2: 0.0534, m: 0.0217 and $n = 500$: β_1: 0.0375; β_2: 0.0287, m: 0.0145

Table 5.4: Simulation Model 2: Average ratio of RE($\hat{m}$) based on $\boldsymbol{\lambda}_{CV}$ and $\boldsymbol{\lambda}_{opt}$ for the P-spline estimates using transformation methods 1 (M1) and 2 (M2). M1$_{CV}$: M1 with cross-validation transformation, M1$_{MV}$: M1 with minimal-variability transformation, M$_K$: M1 with transformation by Koul et al. (1981) using $\gamma = -1$. n is the sample size, PC is the percentage of censoring.

	$n = 250$				$n = 500$			
PC	M1$_{CV}$	M1$_{MV}$	M1$_K$	M2	M1$_{CV}$	M1$_{MV}$	M1$_K$	M2
10	1.2126	1.2152	1.3705	1.1529	1.2028	1.1959	1.3234	1.1462
30	1.2349	1.2281	1.4376	1.1342	1.2563	1.2560	1.3614	1.1620
50	1.1688	1.1589	1.2220	1.0875	1.1686	1.1354	1.1328	1.0705



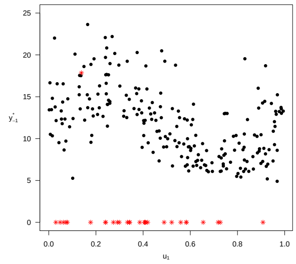
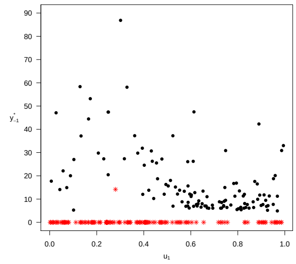
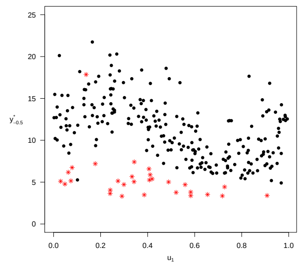
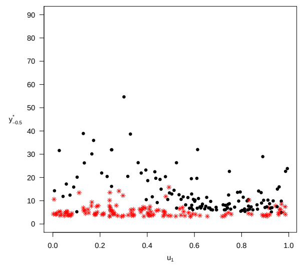
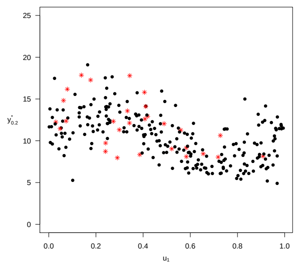
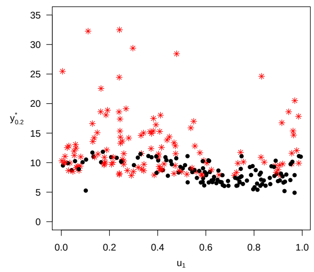
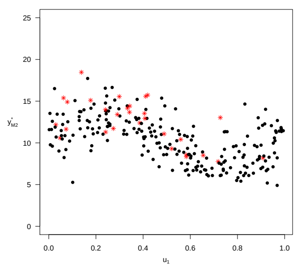
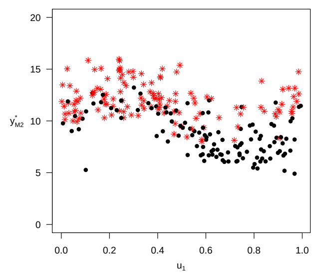
(a) **(b)**

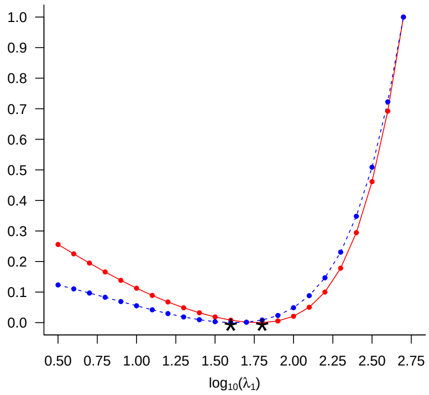
Figure 5.1: Simulation Model 2: Pointwise confidence bands for the P-spline estimates of the regression parameter functions (black, solid) (a) β_1 and (b) β_2 obtained with method 1 ($\gamma = -1$) (green, dashed dotted), method 1 and γ_{MV} (blue, dotted) and method 2 (red, dashed) for $n = 500$ and PC $= 30\%$.

5.8 Real data example: Addict data

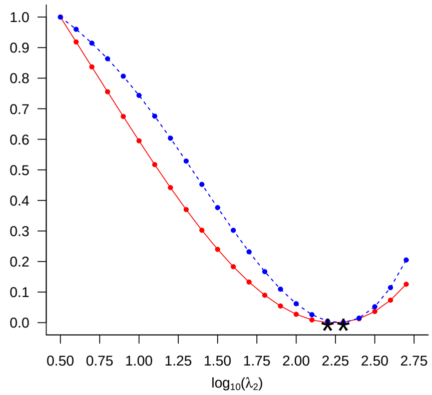
In a study by Caplehorn and Bell (1991), data were collected on a cohort of 238 heroin addicts, who entered maintenance programs between February 1986 and August 1987, to study retention of patients in methadone treatment. All patients had been referred to

Table 5.5: Simulation Model 2: Responses transformed with method 1 (M1) for different choices of γ and method 2 (M2) for PC = 10% and PC = 50% for $n = 250$. Uncensored observations are indicated by black dots, censored observations are indicated by red asterisks.

	10%	50%
$M1(\gamma = -1)$		
$M1(\gamma = -0.5)$		
$M1(\gamma = 0.2)$		
$M2$		



(a)



(b)

Figure 5.2: Simulation Model 2: (a) CV (red,solid) and relative regression error (blue, dashed) curves for $\lambda_1 \in 10^{\{0.5,0.6,\ldots,2.6\}}$ and for λ_2 minimizing CV resp. relative error. (b) CV (red, solid) and relative regression error (blue, dashed) curves for $\lambda_2 \in 10^{\{0.5,0.6,\ldots,2.6\}}$ and for λ_1 minimizing CV resp. relative error. The black asterisk indicates the minimal value. The curves are based on one simulated data set of size $n = 500$ and PC =30% using method 1 with MV-transformation parameter.

one of two methadone treatment clinics for maintenance. Methadone is a drug similar to heroin which prevents or reduces withdrawal symptoms when a patient stays off heroin. Patients detoxifying from methadone maintenance soon return to illicit opiate abuse, and methadone is only beneficial to addicts in treatment. The main objective of the study was to investigate the effectiveness of treatment programs based on the time an addict spends in a clinic, the larger this duration time the more effective the therapy is. The response is the duration time T (in days) of heroin addicts from entry to a clinic until departure or end of study period; 150 out of the 238 patients left the clinic during the study period, the remaining 88 patients still in the clinic at the end of the study period are censored cases. We focus on the effect of clinic (C, 1= clinic 1, 0 = clinic 2) and a history of imprisonment (P, 1= yes, 0= no) on the time remaining on methadone treatment in a VCM where the coefficients vary with the maximum methadone dosage (M, in mg/day), i.e.

$$\mathbb{E}(T|M, C, P) = \beta_1(M) + \beta_2(M) \times C + \beta_3(M) \times P.$$

Table 5.6: Simulation Model 3: Average relative error for the estimates of the functions (F) β_0, β_1, m and σ^2 obtained with the P-spline estimator using method 1 with minimal-variability transformation (M1) and method 2 (M2). n is the sample size, PC is the percentage of censoring.

n :			250			500		
PC:			10%	30%	50%	10%	30%	50%
α	M	F						
1	M1	β_1	0.1984	0.3393	0.5431	0.1392	0.2731	0.3891
		β_2	0.1371	0.2502	0.4077	0.0995	0.2021	0.2868
		m	0.0411	0.0781	0.1328	0.0306	0.0614	0.0956
	M2	β_1	0.1494	0.2294	0.3778	0.1024	0.1780	0.2735
		β_2	0.0983	0.1769	0.3048	0.0694	0.1399	0.2178
		m	0.0309	0.0580	0.1063	0.0228	0.0456	0.0775
2	M1	β_1	0.2274	0.3647	0.5532	0.1624	0.2787	0.4053
		β_2	0.1538	0.2682	0.4139	0.1127	0.2045	0.3011
		m	0.0468	0.0847	0.1371	0.0355	0.0642	0.1012
	M2	β_1	0.1899	0.2591	0.3880	0.1336	0.1939	0.2811
		β_2	0.1224	0.1954	0.3096	0.0874	0.1463	0.2255
		m	0.0383	0.0655	0.1109	0.0290	0.0491	0.0819
1	M1	σ^2	0.2006	0.3398	0.7071	0.1594	0.2896	0.4480
2	M1	σ^2	0.2158	0.3186	0.4803	0.1674	0.2623	0.3658

We present results for a homoscedastic model based on method 2 only, since method 2 outperformed method 1 in our simulation study and since similar results were obtained with a hetereoscedastic model. We smooth the coefficients by P-splines of degree 3 on 15 equidistant knots with a second order difference penalty. The initial estimate for the regression coefficients is obtained using the first method and an MV-transformation parameter ($\gamma_{MV} = -0.2$). The smoothing parameters ($\lambda_1 = 50, \lambda_2 = 250$ and $\lambda_3 = 100$) were selected by cross-validation on a logarithmic scale. Figure 5.3 presents the resulting estimated mean survival time obtained with transformation method 2 . Only in the second clinic doses above 80 mg/day were given to the patients, however our model reveals that these doses no longer result in larger duration times. This finding could not be obtained if a linear term was considered for the methadone effect. For small methadone doses, the estimated mean survival time is similar for all patients but when the dosage increases, the second clinic tends to do a better job in retaining its patients under treatment. Figure 5.3 also shows that the length of time in treatment is shorter for patients with a history of imprisonment.

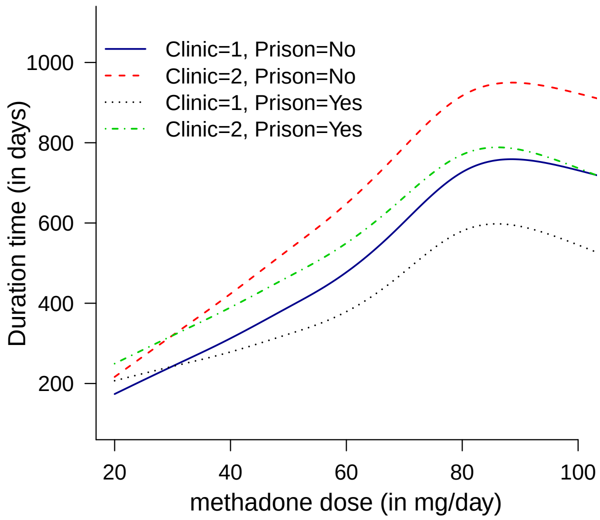


Figure 5.3: Addict data. Fitted P-spline regression function with method 2 using method 1 with $\gamma_{MV} = -0.2$ and $\lambda_{1,CV} = 50, \lambda_{2,CV} = 250, \lambda_{3,CV} = 100.$

Chapter 6

Avenues for future research

The algorithms in Part I for modeling a relationship under a monotonicity constraint all depend on the behavior of a cube-root n statistic. Chapters 2 and 3 focus on the MLE of the distribution function under current status data. The LSE of the monotone link function determines the behavior of the regression parameters in the single index model of Chapter 4. Extensions of our techniques to other cube-root n statistics can be investigated. In this Chapter, we therefore outline some possible avenues for future research that are related to Part I and Part II of this thesis.

The most prominent question that arises with this research concerns the asymptotic behavior of the MLE in the current status linear regression model and the LSE of the monotone single index model. In our attempts to understand the theoretical properties of these estimators, we were always confronted with difficulties that arise when the non-differentiable cube-root n estimator for the nuisance function appears in the optimization criterion for an M-estimator. These problems for the maximization approach were circumvented in the score approach of Chapter 2 and Chapter 4. Simulation results also gave no clear insight in the convergence rate of the estimators. Some experiments were in favor of the parametric $\sqrt{n}$-rate whereas other trials suggested a slower convergence rate. Nevertheless, it remains an interesting topic to understand the behavior of the classical MLE for the current status linear regression model and the LSE for the monotone single index model.

The behavior of the MLE of the distribution function under current status data has been investigated extensively. A lot less is known for the more challenging interval censoring type 2 model. In this setting, one observes an inspection interval (T, U) together with the information whether or not the event time of interest Y lies before the first inspection time T, inside the interval (T, U) or after the second inspection time U. The MLE of the distribution function of Y is defined by the maximizer of the log likelihood given by

$$\ell(F) = \sum_{i=1}^{n} \left[\Delta_{i1} \log F(T_i) + \Delta_{i2} \log\{F(U_i) - F(T_i)\} + \Delta_{i3} \log\{1 - F(U_i)\}\right],$$

where $\Delta_{i1} = 1_{\{Y_i \le T_i\}}, \Delta_{i2} = 1_{\{T_i < Y_i < U_i\}}$ and $\Delta_{i3} = 1 - \Delta_{i1} - \Delta_{i2}$ for $1 \le i \le n$.
A distinction can be made between the separated and the non-separated case. In the separated case one assumes that T is strictly smaller than U. For this case the asymptotic limit distribution of the MLE of F_0 was derived in Groeneboom (1996). On the other hand, in the non-separated case, the observation intervals can be arbitrarily small and the asymptotic distribution of the MLE has only been conjectured without a complete proof in Groeneboom and Wellner (1992). A transition from the convergence rate $n^{-1/3}$ for the separated case to the rate $(n \log n)^{-1/3}$ for the (conjectured) non-separated case quantifies the small gain in performance when observation intervals (T_i, U_i) are allowed to have arbitrarily short lengths. So far, this conjecture is still not proved and extending our results for the non-separated case will be more challenging than for the separated case. Nevertheless, our algorithm to estimate a regression parameter in a linear regression model under censored data can be implemented for interval censoring type 2 and would lead to the study of the score function

$$\sum_{(T_i,U_i)\in A_{n,\boldsymbol{\alpha},\varepsilon}} \boldsymbol{X}_i \left\{ \frac{\Delta_{i1}\hat{f}_{n,\boldsymbol{\alpha}}(T_i - \boldsymbol{\alpha}^T\boldsymbol{X}_i)}{\hat{F}_{n,\boldsymbol{\alpha}}(T_i - \boldsymbol{\alpha}^T\boldsymbol{X}_i)} + \frac{\Delta_{i2}\{\hat{f}_{n,\boldsymbol{\alpha}}(U_i - \boldsymbol{\alpha}^T\boldsymbol{X}_i) - \hat{f}_{n,\boldsymbol{\alpha}}(T_i - \boldsymbol{\alpha}^T\boldsymbol{X}_i)\}}{\hat{F}_{n,\boldsymbol{\alpha}}(U_i - \boldsymbol{\alpha}^T\boldsymbol{X}_i) - \hat{F}_{n,\boldsymbol{\alpha}}(T_i - \boldsymbol{\alpha}^T\boldsymbol{X}_i)} - \frac{\Delta_{i3}\hat{f}_{n,\boldsymbol{\alpha}}(U_i - \boldsymbol{\alpha}^T\boldsymbol{X}_i)}{1 - \hat{F}_{n,\boldsymbol{\alpha}}(U_i - \boldsymbol{\alpha}^T\boldsymbol{X}_i)} \right\},$$

where $\hat{F}_{n,\boldsymbol{\alpha}}$ is the MLE of the distribution function for fixed $\boldsymbol{\alpha}$, $\hat{f}_{n,\boldsymbol{\alpha}}$ is an estimate of the density, and $A_{n,\boldsymbol{\alpha},\varepsilon}$ is a well-defined truncation interval to avoid division by zero in the above score expression. Deriving the behavior of score estimators of the regression parameter in a linear regression model under interval censored data is worth studying in further research. For the density estimate $\hat{f}_{n,\boldsymbol{\alpha}}$, a kernel-based smooth version of the MLE $\hat{F}_{n,\boldsymbol{\alpha}}$ can be considered in the same way as the smoothing approaches introduced in Chapter 2 and Chapter 3. This immediately points out a possible extension of the results presented in Chapter 3, namely investigating the behavior of (smooth) bootstrap procedures for the distribution function under interval censoring type 2. The limiting distribution of the Smoothed MLE for the separated case, obtained in the same way as the SMLE in Chapter 3 is conjectured in Groeneboom and Ketelaars (2011). Proving that this conjecture holds is the first step towards developing the results of Chapter 3. The quality of the bootstrap procedures for interval censoring type 2 is still unexplored.

Furthermore, the MLE for the separated interval censoring type 2 model is not the only cube-root n statistic that can be explored in the setting of Chapter 3. In monotone density estimation, one could define a smoothed version of the Grenander estimator (introduced in Section 1). The isotonic link function in the monotone single index model of Chapter 4 can also be extended similarly. Deriving the limiting distribution of these smoothed monotone estimators or investigating applications of the bootstrap in order to obtain confidence intervals for the corresponding monotone functions are interesting avenues for further research. Moreover, the asymptotic behavior of the smooth derivative estimate introduced in Section 4.3.2 is still unknown. We expect that the methods introduced in Chapter 3 behave similarly in these monotone density and single index model settings.

Another interesting extension of the research in Chapter 3 is the construction of confidence bands for the distribution function instead of the currently proposed pointwise CIs. Note that our main results do not imply that

$$\mathbb{E}\left\{\sup_{t\in[0,M]} n^{1/3}\left|\hat{F}_n^*(t) - F_0(t)\right|\,\middle|\, Z_1,\ldots,Z_n\right\} = O_p(1).$$

A bound on $\sup_{t\in[0,M]} n^{1/3}\left|\hat{F}_n^*(t) - F_0(t)\right|\Big|$ would be needed for confidence bands instead of our pointwise CIs. Such bounds would, without a doubt, contain logarithmic factors. The idea is that the process $t \mapsto n^{1/3}\{\hat{F}_n^*(t) - F_0(t)\}$ will fall apart into asymptotically independent pieces, and that we therefore expect Gumbel-type distributions to enter, via the maximum of independent random variables. The theory for this still has to be developed, however. What strike us in the present simulation studies is how comparatively well the global behavior of our pointwise CIs still was, indicating that the extra logarithmic factors do not have such a very large impact.

Instead of kernel smoothing techniques, we used P-spline smoothing in Part II. It is therefore interesting to develop theory for smoothed versions of the monotone estimators of Part I using splines and to contrast the performance between kernel and spline estimators of the monotone functions.

For data subject to right censoring, the synthetic data approach discussed in Chapter 5 is considered for heteroscedastic models. We did not consider variance-based weighting in the estimation of the mean regression curve. Although it is common practice to use weighted least squares when heterogeneity is present in the data (e.g. for non-censored data, Shen et al. (2016) use reweighting for heteroscedastic VCMs), Antoniadis et al. (2012) show the good performance of P-spline estimators in VCMs with non-censored data even if the heteroscedasticity is ignored in the estimation process. How to bring in variance-based reweighting in the estimation process and studying the impact of reweighting on the quality of the P-spline estimators in heteroscedastic VCMs is a challenging open problem.

Finally we note that the VCM of Part II for right censored observations can be considered for interval censored observations as well. For the data transformations that require estimation of the underlying error distribution, the cube-root n MLE can be considered. The Kaplan-Meier estimator (Kaplan and Meier, 1958), which is the MLE for right censored data, converges however at a faster $\sqrt{n}$-rate because of the fact that one has actual observations in addition to the censored ones. Since all observations are censored in the current status model and the interval censoring type 2 model, additional complexities are to be expected. The method proposed by Buckley and James (1979), which is the second transformation method in Chapter 5, has been implemented for interval censored data by Rabinowitz et al. (1995) but more work is still needed to understand the theoretical properties of this estimator. The construction of model-independent data transformations has also been considered in Zheng (2008) for a linear regression model under interval censored data. Finding, for interval censored data, appropriate transformations for inference in VCMs is an unexplored interesting open question.

Appendices

Appendix A

Current status linear regression - Appendix

We give the proofs of the results stated in Chapter 2 on the asymptotic behavior of the simple score estimator $\hat{\boldsymbol{\alpha}}_{1n}$ (SSE), the efficient score estimator $\hat{\boldsymbol{\alpha}}_{2n}$ (ESE) and the plug-in estimator $\hat{\boldsymbol{\alpha}}_{3n}$ for the regression parameter $\boldsymbol{\alpha}_0$ in the current status linear regression model (1.1.3). The proofs for each method are given in Sections A.2, A.3 and A.4. To simplify the notations we drop the index $j = 1, 2, 3$ in $\boldsymbol{\alpha}_{jn}$, representing the different techniques.

Entropy results are used in our proofs. Before we prove the results we first give some definitions and an equicontinuity lemma needed in the proofs.

Consider a class of functions $\mathcal{F}$ on $\mathcal{R}$ and let $L_2(Q)$ be the L_2-norm defined by a probability measure Q on $\mathcal{R}$, i.e. for $g \in \mathcal{F}$,

$$\|g\|_{L_2} = \int g^2 dQ.$$

For any probability measure Q on $\mathcal{R}$ let $N_B(\zeta, \mathcal{F}, L_2(Q))$ be the minimal number N for which there exists pairs of functions $\{[g_j^L, g_j^U], 1 \le j \le N\}$ such that $\|g_j^U - g_j^L\|_{L_2} \le \zeta$ for all $1 \le j \le N$ and such that for each $g \in \mathcal{F}$ there is a $j \in \{1, \ldots, N\}$ such that $g_j^L \le g \le g_j^U$. The $\zeta-$entropy with bracketing of $\mathcal{F}$ (for the $L_2(Q)-$distance) is defined as $H_B(\zeta, \mathcal{F}, L_2(Q)) = \log(N_B(\zeta, \mathcal{F}, L_2(Q)))$.

Lemma A.0.1 (Equicontinuity Lemma, Theorem 5.12, p.77 in van de Geer (2000))**.** *Let $\mathcal{F}$ be a fixed class of functions with envelope F in $L_2(P) = \{f : \int f^2 dP < \infty\}$. Suppose*

that
$$\int_0^1 H_B^{1/2}(u, \mathcal{F}, L_2(P))\, du \le \infty,$$
where H_B is the entropy with bracketing of $\mathcal{F}$ for the L_2-norm. Then, for all $\eta > 0$ there exists a $\delta > 0$ such that
$$\limsup_{n\to\infty} P\left(sup_{[\delta]} |\sqrt{n} \int (f-g) d(\mathbb{P}_n - P_0)| > \eta \right) < \eta,$$
where
$$[\delta] = \{(f,g) : \|f - g\| \le \delta\}.$$

A.1 The maximum likelihood estimator $\hat{F}_{n,\alpha}$

In this section we prove Lemma 2.3.1. We first prove in Lemma A.1.1 some entropy bounds needed in the proofs.

Lemma A.1.1. *Let*
$$\mathcal{F} = \{(\boldsymbol{x}, t) \mapsto F(t - \boldsymbol{\alpha}^T\boldsymbol{x}) : F \in \mathcal{F}_0, \boldsymbol{\alpha} \in \Theta\},$$
where $\mathcal{F}_0$ is the set of subdistribution functions on $[a,b]$, where $[a,b]$ contains all values $t - \boldsymbol{\alpha}^T\boldsymbol{x}$ for $\boldsymbol{\alpha} \in \Theta$, and $(\boldsymbol{x}, t)$ in the compact neighborhood over which we let them vary. Then,
$$\sup_{\varepsilon > 0} \varepsilon H_B\left(\varepsilon, \mathcal{F}, L_2(P_0)\right) = O(1),$$
Furthermore, let
$$\mathcal{G} = \{(\boldsymbol{x}, t) \mapsto g(t - \boldsymbol{\alpha}^T\boldsymbol{x}) : g \in \mathcal{G}_0, \boldsymbol{\alpha} \in \Theta\},$$
where $\mathcal{G}_0$ is a set functions of uniformly bounded variation, then
$$\sup_{\varepsilon > 0} \varepsilon H_B\left(\varepsilon, \mathcal{G}, L_2(P_0)\right) = O(1). \tag{A.1.1}$$

Proof. We only prove the result for the class $\mathcal{F}$ since the proof for the class $\mathcal{G}$ can be obtained similarly.
Fix $\varepsilon > 0$. We first note that Θ can be covered by N neighborhoods with diameter at most ε^2 where N is of order ε^{-2d}. Let $\{\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_N\}$ denote elements of each of these neighborhoods. Consider an ε-bracket $[F_j^L, F_j^U], 1 \le j \le N'$ covering the class $\mathcal{F}_0$ such that
$$\left\{ \int \{F_j^U(u) - F_j^L(u)\}^2 f_{T - \boldsymbol{\alpha}^T \boldsymbol{X}}(u)\, du \right\}^{1/2} < \varepsilon.$$

for $1 \le j \le N'$. The existence of such an ε-net is assured by the fact that $f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}$ is bounded above (uniformly in $\boldsymbol{\alpha}$). The number N' is of order $\exp(C/\varepsilon)$ for some constant C (See e.g. van de Geer (2000), p. 18). Let $\boldsymbol{\alpha}_j$ be chosen such that $\|\boldsymbol{\alpha}_j - \boldsymbol{\alpha}\| < \varepsilon^2$, where $\|\cdot\|$ denotes the Euclidean norm. Then:

$$t - \boldsymbol{\alpha}_j^T\boldsymbol{x} - \varepsilon^2 R \le t - \boldsymbol{\alpha}^T\boldsymbol{x} = t - \boldsymbol{\alpha}_j^T\boldsymbol{x} - (\boldsymbol{\alpha} - \boldsymbol{\alpha}_j)^T\boldsymbol{x} \le t - \boldsymbol{\alpha}_j^T\boldsymbol{x} + \varepsilon^2 R,$$

where R is the maximum of the values $\|\boldsymbol{x}\|$. This implies that for each $F \in \mathcal{F}_0$ and $\boldsymbol{\alpha} \in \Theta$,

$$F_i^L(t - \boldsymbol{\alpha}_j^T\boldsymbol{x} - \varepsilon^2 R) \le F(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \le F_i^U(t - \boldsymbol{\alpha}_j^T\boldsymbol{x} + \varepsilon^2 R),$$

for some $1 \le i \le N'$ and $1 \le j \le N$. The result of Lemma A.1.1 follows if we can show that

$$\left\{\int \{F_i^U(t - \boldsymbol{\alpha}_j^T\boldsymbol{x} + \varepsilon^2 R) - F_i^L(t - \boldsymbol{\alpha}_j^T\boldsymbol{x} - \varepsilon^2 R)\}^2 dG(\boldsymbol{x}, t)\right\}^{1/2} \lesssim \varepsilon. \tag{A.1.2}$$

By the triangle inequality we get that the left-hand side of the above equation is bounded by

$$\begin{aligned}
&\left\{\int \{F(t - \boldsymbol{\alpha}_j^T\boldsymbol{x} - \varepsilon^2 R) - F_i^L(t - \boldsymbol{\alpha}_j^T\boldsymbol{x} - \varepsilon^2 R)\}^2\, dG(\boldsymbol{x}, t)\right\}^{1/2}\\
&\quad + \left\{\int \{F_i^U(t - \boldsymbol{\alpha}_j^T\boldsymbol{x} + \varepsilon^2 R) - F(t - \boldsymbol{\alpha}_j^T\boldsymbol{x} + \varepsilon^2 R)\}^2\, dG(\boldsymbol{x}, t)\right\}^{1/2}\\
&\quad + \left\{\int \{F(t - \boldsymbol{\alpha}_j^T\boldsymbol{x} + \varepsilon^2 R) - F(t - \boldsymbol{\alpha}_j^T\boldsymbol{x} - \varepsilon^2 R)\}^2\, dG(\boldsymbol{x}, t)\right\}^{1/2}\\
&\lesssim \varepsilon + \left\{\int \{F(u + \varepsilon^2 R) - F(u - \varepsilon^2 R)\}^2 f_{T-\boldsymbol{\alpha}_j^T\boldsymbol{x}}(u)\right\}^{1/2}.
\end{aligned}$$

Let $u_0 = a - \varepsilon^2 R < u_1, \cdots < u_m = b + \varepsilon^2 R$, be points such that $u_k - u_{k-1} = \varepsilon^2$, $1 \le k \le m-1$, $u_m - u_{m-1} \le \varepsilon^2$. Then:

$$\begin{aligned}
&\int \{F(u + \varepsilon^2 R) - F(u - \varepsilon^2 R)\}^2 f_{T-\boldsymbol{\alpha}_j^T\boldsymbol{X}}(u)\, du\\
&\le \int \{F(u + \varepsilon^2 R) - F(u - \varepsilon^2 R)\} f_{T-\boldsymbol{\alpha}_j^T\boldsymbol{X}}(u)\, du\\
&\le M \int \{F(u + \varepsilon^2 R) - F(u - \varepsilon^2 R)\}\, du\\
&= M \int_{a+\varepsilon^2 R}^{b+\varepsilon^2 R} F(u)\, du - M \int_{a-\varepsilon^2 R}^{b-\varepsilon^2 R} F(u)\, du\\
&\le M \int_{a-\varepsilon^2 R}^{a+\varepsilon^2 R} F(u)\, du + M \int_{b-\varepsilon^2 R}^{b+\varepsilon^2 R} F(u)\, du \lesssim \varepsilon^2,
\end{aligned}$$

where M is an upper bound for $f_{T-\boldsymbol{\alpha}_j^T\boldsymbol{X}}$, and where we extend the function F by a constant value outside $[a,b]$. This completes the proof of (A.1.2) since we have shown that there exist positive constants A_1, A_2, A_3 and C such that

$$\begin{aligned} H_B\left(\varepsilon, \mathcal{F}, L_2(P_0)\right) &\le \log N + \log N' \le d\log(A_1/\varepsilon^2) + A_2\log(\exp(C/\varepsilon)) \\ &\le A_3/\varepsilon \\ &= O(\log(1/\varepsilon)) + O(1/\varepsilon) = O(1/\varepsilon), \qquad \varepsilon \downarrow 0. \end{aligned}$$

□

Proof of Lemma 2.3.1. Let h denote the Hellinger distance on the class of densities $\mathcal{P}$ defined by

$$\mathcal{P} = \left\{p_{\boldsymbol{\alpha},F}(\boldsymbol{x},t,\delta) = \delta F(t-\boldsymbol{\alpha}^T\boldsymbol{x}) + (1-\delta)\{1-F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\} : F \in \mathcal{F}_0, \boldsymbol{\alpha} \in \Theta\right\},$$

w.r.t. the product of counting measure on $\{0,1\}$ and the measure dG of $(\boldsymbol{X},T)$, where $\mathcal{F}_0$ is the class of right-continuous subdistribution functions.

We have (see e.g. the "basic inequality" Lemma 4.5, p. 51 of van de Geer (2000)):

$$h^2\left(p_{\boldsymbol{\alpha},\hat{F}_{n,\boldsymbol{\alpha}}}, p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}\right) \le \int \frac{2p_{\boldsymbol{\alpha},\hat{F}_{n,\boldsymbol{\alpha}}}}{p_{\boldsymbol{\alpha},\hat{F}_{n,\boldsymbol{\alpha}}} + p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}}\, d\left(\mathbb{P}_n - P_0\right),$$

where we use the convexity of the set of densities of this type for (temporarily) fixed $\boldsymbol{\alpha}$.

Hence we get, for $\epsilon \in (0,1]$:

$$
\begin{aligned}
&\mathbb{P}\left\{\sup_{\boldsymbol{\alpha}\in\Theta} h\left(p_{\boldsymbol{\alpha},\hat F_{n,\boldsymbol{\alpha}}}, p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}\right) \ge \epsilon\right\}\\
&= \mathbb{P}\Bigg\{ \sup_{\substack{\boldsymbol{\alpha}\in\Theta,\\ h\left(p_{\boldsymbol{\alpha},\hat F_{n,\boldsymbol{\alpha}}}, p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}\right)\ge\epsilon}} \Bigg\{\int\Bigg\{\frac{2p_{\boldsymbol{\alpha},\hat F_{n,\boldsymbol{\alpha}}}}{p_{\boldsymbol{\alpha},\hat F_{n,\boldsymbol{\alpha}}}+p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}}-1\Bigg\}\,d\left(\mathbb{P}_n-P_0\right)\\
&\qquad\qquad - h^2\left(p_{\boldsymbol{\alpha},\hat F_{n,\boldsymbol{\alpha}}}, p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}\right)\Bigg\} \ge 0, \quad \sup_{\boldsymbol{\alpha}\in\Theta} h\big(p_{\boldsymbol{\alpha},\hat F_{n,\boldsymbol{\alpha}}},\, p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}\big) \ge \epsilon\Bigg\}\\
&\le \mathbb{P}\Bigg\{ \sup_{\substack{\boldsymbol{\alpha}\in\Theta,\\ F\in\mathcal{F}_0,\, h\left(p_{\boldsymbol{\alpha},F}, p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}\right)\ge\epsilon}} \Bigg\{\int\Bigg\{\frac{2p_{\boldsymbol{\alpha},F}}{p_{\boldsymbol{\alpha},F}+p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}}-1\Bigg\}\,d\left(\mathbb{P}_n-P_0\right)\\
&\qquad\qquad - h^2\big(p_{\boldsymbol{\alpha},F},\, p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}\big)\Bigg\} \ge 0\Bigg\}\\
&\le \sum_{s=0}^{\infty}\mathbb{P}\Bigg\{ \sup_{\substack{\boldsymbol{\alpha}\in\Theta,\, F\in\mathcal{F}_0,\\ 2^s\epsilon\le h\left(p_{\boldsymbol{\alpha},F}, p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}\right)\le 2^{s+1}\epsilon}} \sqrt{n}\int\left\{\frac{2p_{\boldsymbol{\alpha},F}}{p_{\boldsymbol{\alpha},F}+p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}}-1\right\}\,d\left(\mathbb{P}_n-P_0\right)\\
&\qquad\qquad \ge \sqrt{n}\,2^{2s}\epsilon^2\Bigg\},
\end{aligned}
$$

We can now use Theorem 5.13 in van de Geer (2000), taking $\epsilon = Mn^{-1/3}, \alpha = 1, \beta = 0$ and $T = \sqrt{n}\,2^{2s}\epsilon^2 = M2^{2s}n^{-1/6}$, together with Lemma A.1.1 for the entropy of the set of densities to conclude:

$$
\begin{aligned}
&\sum_{s=0}^{\infty}\mathbb{P}\Bigg\{ \sup_{\substack{\boldsymbol{\alpha}\in\Theta,\, F\in\mathcal{F}_0,\\ 2^s\epsilon\le h\left(p_{\boldsymbol{\alpha},F}, p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}\right)\le 2^{s+1}\epsilon}} \sqrt{n}\int\left\{\frac{2p_{\boldsymbol{\alpha},F}}{p_{\boldsymbol{\alpha},F}+p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}}-1\right\}\,d\left(\mathbb{P}_n-P_0\right)\\
&\qquad\qquad \ge \sqrt{n}\,2^{2s}\epsilon^2\Bigg\},\\
&\quad \le \sum_{s=0}^{\infty} c_1\exp(-c_2M2^{2s})
\end{aligned}
$$

for constants $c_1, c_2 > 0$. Since the sum can be made arbitrarily small for M sufficiently large, we find:

$$
\sup_{\boldsymbol{\alpha}\in\Theta} h\left(p_{\boldsymbol{\alpha},\hat F_{n,\boldsymbol{\alpha}}}, p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}\right) = O_p\left(n^{-1/3}\right).
$$

We have:

$$h\left(p_{\boldsymbol{\alpha},\hat F_{n,\boldsymbol{\alpha}}},p_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}\right)^2=\tfrac12\int\left\{p^{1/2}_{\boldsymbol{\alpha},\hat F_{n,\boldsymbol{\alpha}}}(\boldsymbol{x},t,1)-p^{1/2}_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}(\boldsymbol{x},t,1)\right\}^2dG(\boldsymbol{x},t)$$
$$+\tfrac12\int\left\{p^{1/2}_{\boldsymbol{\alpha},\hat F_{n,\boldsymbol{\alpha}}}(\boldsymbol{x},t,0)-p^{1/2}_{\boldsymbol{\alpha},F_{\boldsymbol{\alpha}}}(\boldsymbol{x},t,0)\right\}^2dG(\boldsymbol{x},t)$$
$$=\tfrac12\int\left\{\hat F_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})^{1/2}-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})^{1/2}\right\}^2dG(\boldsymbol{x},t)$$
$$+\tfrac12\int\left\{\left(1-\hat F_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right)^{1/2}-\left(1-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right)^{1/2}\right\}^2dG(\boldsymbol{x},t),$$

and

$$\int\left\{\hat F_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\}^2dG(\boldsymbol{x},t)$$
$$=\int\left\{\hat F_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})^{1/2}-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})^{1/2}\right\}^2$$
$$\cdot\left\{\hat F_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})^{1/2}+F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})^{1/2}\right\}^2dG(\boldsymbol{x},t)$$
$$\le 4\int\left\{\hat F_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})^{1/2}-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})^{1/2}\right\}^2dG(\boldsymbol{x},t)$$
$$\le 8h\left(p_{\hat F_{n,\boldsymbol{\alpha}}},p_{F_{\boldsymbol{\alpha}}}\right)^2.$$

So we find:

$$\sup_{\boldsymbol{\alpha}\in\Theta}\int\left\{\hat F_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\}^2dG(\boldsymbol{x},t)=O_p\left(n^{-2/3}\right).$$

□

A.2 Asymptotic behavior of the SSE

This section contains the proof of Theorem 2.4.1 stated in Section 2.4.1 of the thesis. The proof is decomposed into three parts: (a) proof of existence of a root of $\psi_{1\epsilon,n}$, (b) proof of consistency of $\hat{\boldsymbol{\alpha}}_n$ and (c) proof of asymptotic normality of $\sqrt{n}(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)$. We first prove the properties given in Lemma 2.4.1 of the population version of the statistic $\psi_{1\epsilon,n}$ defined by

$$\psi_{1\epsilon}(\boldsymbol{\alpha})=\int_{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]}\boldsymbol{x}\{\delta-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\}\,dP_0(\boldsymbol{x},t,\delta)$$
$$=\int_{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]}\boldsymbol{x}\{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\}\,dG(\boldsymbol{x},t). \tag{A.2.1}$$

Proof of Lemma 2.4.1 . We first note that

$$\psi_{1\epsilon}(\boldsymbol{\alpha}_0) = \int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\Big\{\mathbb{E}\{\Delta|(\boldsymbol{X},T)=(\boldsymbol{x},t)\} - F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\Big\}\,dG(\boldsymbol{x},t)$$
$$= \mathbf{0},$$

since $\mathbb{E}\{\Delta|(\boldsymbol{X},T)=(\boldsymbol{x},t)\} = F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})$. We next continue by showing result (i). Since

$$\mathbb{E}\left(\Delta|T-\boldsymbol{\alpha}^T\boldsymbol{X}=t-\boldsymbol{\alpha}^T\boldsymbol{x}\right) = F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}),$$

we get:

$$\begin{aligned}
&\mathbb{E}_{\epsilon,\boldsymbol{\alpha}}\left[\mathsf{Cov}\left(\Delta,\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}\right)\right]\\
&\overset{\text{def}}{=}\int_{F_{\boldsymbol{\alpha}}(u)\in[\epsilon,1-\epsilon]} \mathsf{Cov}\left(\Delta,\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}=u\right) f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(u)\,du\\
&=\int_{F_{\boldsymbol{\alpha}}(u)\in[\epsilon,1-\epsilon]} \mathsf{Cov}\left\{\boldsymbol{X},F_0(u+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{X})\;\Big|\;T-\boldsymbol{\alpha}^T\boldsymbol{X}=u\right\} f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(u)\,du\\
&=\int_{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\Big\{F_0(t-\boldsymbol{\alpha}^T\boldsymbol{x}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{x}) - F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\Big\}\,dG(\boldsymbol{x},t)\\
&=\int_{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\left\{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\}\,dG(\boldsymbol{x},t) = \psi_{1\epsilon}(\boldsymbol{\alpha}).
\end{aligned}$$

For the second result (ii), we write:

$$\begin{aligned}
&(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\mathsf{Cov}\left(\Delta,\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}=u\right)\\
&\quad=\mathsf{Cov}\left(F_0(T-\boldsymbol{\alpha}^T\boldsymbol{X}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{X}),(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}=u\right),
\end{aligned}$$

which is positive for all $\boldsymbol{\alpha}$, following from the fact that F_0 is an increasing function. Indeed, using Fubini's theorem, one can prove that for any random variables X and Y such that XY, X and Y are integrable, we have:

$$\begin{aligned}
\mathsf{Cov}\{X,Y\} &= EXY - EXEY\\
&= \int\{\mathbb{P}(X\geq s, Y\geq t) - \mathbb{P}(X\geq s)\mathbb{P}(Y\geq t)\}\,ds\,dt.
\end{aligned}$$

Denote $Z_1 = (\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{X}$ and $Z_2 = F_0(u+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{X}) = F_0(u+Z_1)$. For simplicity of notation we no longer write the condition $T-\boldsymbol{\alpha}^T\boldsymbol{X}=u$ but note that the results below hold conditional on $T-\boldsymbol{\alpha}^T\boldsymbol{X}=u$. Using the monotonicity of the function F_0, we have:

$$\begin{aligned}
\mathbb{P}(Z_1\geq z_1, Z_2\geq z_2) &= \mathbb{P}(Z_1\geq \max\{z_1,\tilde{z}_2\})\\
&\geq \mathbb{P}(Z_1\geq \max\{z_1,\tilde{z}_2\})\mathbb{P}(Z_1\geq \min\{z_1,\tilde{z}_2\})\\
&= \mathbb{P}(Z_1\geq z_1)\mathbb{P}(Z_2\geq z_2),
\end{aligned}$$

where

$$\tilde{z}_2 = F_0^{-1}(z_2) - u.$$

We conclude that

$$\begin{aligned}&\mathsf{Cov}\left(F_0(T-\boldsymbol{\alpha}^T\boldsymbol{X}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{X}),(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}=u\right)\\&\quad=\int\{\mathbb{P}(Z_1\ge z_1,Z_2\ge z_2)-\mathbb{P}(Z_1\ge z_1)\mathbb{P}(Z_2\ge z_2)\}\,dz_1\,dz_2\ge 0,\end{aligned}$$

and hence (ii) follows from the assumption that the covariance $\mathsf{Cov}(\boldsymbol{X}, F_0(u+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{X})|T-\boldsymbol{\alpha}^T\boldsymbol{X}=u)$ is not identically zero for u in the region $A_{\epsilon,\boldsymbol{\alpha}}$, for each $\boldsymbol{\alpha}\in\Theta$, implying:

$$\begin{aligned}&\mathbb{E}_{\epsilon,\boldsymbol{\alpha}}\left[\mathsf{Cov}\left(\Delta,\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}\right)\right]\\&=\int_{F_{\boldsymbol{\alpha}}(u)\in[\epsilon,1-\epsilon]}\mathsf{Cov}\left(F_0(T-\boldsymbol{\alpha}^T\boldsymbol{X}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{X}),(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}=u\right)\\&\qquad\qquad\qquad\qquad\qquad\qquad\cdot f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(u)du\\&\ge 0.\end{aligned}$$

[Uniqueness of $\boldsymbol{\alpha}_0$:]

We next show that $\boldsymbol{\alpha}_0$ is the only value $\boldsymbol{\alpha}_*\in\Theta$ such that $\mathbb{E}_{\epsilon,\boldsymbol{\alpha}}[(\boldsymbol{\alpha}-\boldsymbol{\alpha}_*)^T\mathsf{Cov}(\Delta,\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X})]\ge 0$ for all $\boldsymbol{\alpha}\in\Theta$. We start by assuming that, on the contrary, there exists $\boldsymbol{\alpha}_1\ne\boldsymbol{\alpha}_0$ in Θ such that

$$(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\psi_{1\epsilon}(\boldsymbol{\alpha})\ge 0 \qquad \text{and } (\boldsymbol{\alpha}-\boldsymbol{\alpha}_1)^T\psi_{1\epsilon}(\boldsymbol{\alpha})\ge 0 \qquad \text{for all } \boldsymbol{\alpha}\in\Theta,$$

and we consider the point $\tilde{\boldsymbol{\alpha}}\in\Theta$ given by

$$\tilde{\boldsymbol{\alpha}}=\tfrac{1}{2}\{\boldsymbol{\alpha}_0+\boldsymbol{\alpha}_1\}.$$

The existence of the point $\tilde{\boldsymbol{\alpha}}$ is ensured by the convexity of the set Θ. For this point, we have:

$$(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)^T\psi_{1\epsilon}(\tilde{\boldsymbol{\alpha}})=-(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_1)^T\psi_{1\epsilon}(\tilde{\boldsymbol{\alpha}}),$$

which is not possible since both terms should be positive and $\psi_{1\epsilon}(\tilde{\boldsymbol{\alpha}})$ is not equal to zero (since, by the assumption that the covariance $\mathsf{Cov}(\boldsymbol{X},F_0(u+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{X})|T-\boldsymbol{\alpha}^T\boldsymbol{X}=u)$ is not identically zero for u in the region $A_{\epsilon,\boldsymbol{\alpha}}$, $\psi_{1\epsilon}(\boldsymbol{\alpha})$ is only zero at $\boldsymbol{\alpha}=\boldsymbol{\alpha}_0$.)

We now calculate the derivative of $\psi_{1\epsilon}$ at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$. We have:

$$\begin{aligned}
\psi_{1\epsilon}'(\boldsymbol{\alpha}) &= \frac{\partial}{\partial \boldsymbol{\alpha}} \int_{F_{\boldsymbol{\alpha}}^{-1}(\epsilon) \le t - \boldsymbol{\alpha}^T \boldsymbol{x} \le F_{\boldsymbol{\alpha}}^{-1}(1-\epsilon)} \boldsymbol{x}\{\delta - F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\} dP_0(\boldsymbol{x}, t, \delta) \\
&= \frac{\partial}{\partial \boldsymbol{\alpha}} \int_{F_{\boldsymbol{\alpha}}^{-1}(\epsilon) \le t - \boldsymbol{\alpha}^T \boldsymbol{x} \le F_{\boldsymbol{\alpha}}^{-1}(1-\epsilon)} \boldsymbol{x}\{F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x}) - F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\}\, dG(\boldsymbol{x}, t) \\
&= \frac{\partial}{\partial \boldsymbol{\alpha}} \int_{u = F_{\boldsymbol{\alpha}}^{-1}(\epsilon)}^{F_{\boldsymbol{\alpha}}^{-1}(1-\epsilon)} \int \boldsymbol{x}\left\{F_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) - F_{\boldsymbol{\alpha}}(u)\right\} f_{\boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|u) \\
&\qquad\qquad \cdot f_{T - \boldsymbol{\alpha}^T\boldsymbol{X}}(u)\, d\boldsymbol{x}\, du \\
&= \int_{u = F_{\boldsymbol{\alpha}}^{-1}(\epsilon)}^{F_{\boldsymbol{\alpha}}^{-1}(1-\epsilon)} \int \frac{\partial}{\partial \boldsymbol{\alpha}} \{\boldsymbol{x}\left\{F_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) - F_{\boldsymbol{\alpha}}(u)\right\} f_{\boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|u) \\
&\qquad\qquad \cdot f_{T - \boldsymbol{\alpha}^T\boldsymbol{X}}(u)\}\, d\boldsymbol{x}\, du \\
&\quad + \left\{\frac{\partial}{\partial \boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}^{-1}(1-\epsilon)\right\} \int \boldsymbol{x}\left\{F_0(F_{\boldsymbol{\alpha}}^{-1}(1-\epsilon) + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) - (1-\epsilon)\right\} \\
&\qquad\qquad \cdot f_{\boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|F_{\boldsymbol{\alpha}}^{-1}(1-\epsilon)) f_{T - \boldsymbol{\alpha}^T\boldsymbol{X}}(F_{\boldsymbol{\alpha}}^{-1}(1-\epsilon))\, d\boldsymbol{x} \\
&\quad - \left\{\frac{\partial}{\partial \boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}^{-1}(\epsilon)\right\} \int \boldsymbol{x}\left\{F_0(F_{\boldsymbol{\alpha}}^{-1}(\epsilon) + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) - \epsilon\right\} \\
&\qquad\qquad \cdot f_{\boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|F_{\boldsymbol{\alpha}}^{-1}(\epsilon)) f_{T - \boldsymbol{\alpha}^T\boldsymbol{X}}(F_{\boldsymbol{\alpha}}^{-1}(\epsilon))\, d\boldsymbol{x}.
\end{aligned}$$

Note that if $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$, we get:

$$\begin{aligned}
\psi_{1\epsilon}'(\boldsymbol{\alpha}_0) = \int_{F_0^{-1}(\epsilon)}^{F_0^{-1}(1-\epsilon)} \int \frac{\partial}{\partial \boldsymbol{\alpha}} \Big\{ &\boldsymbol{x}\left\{F_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) - F_{\boldsymbol{\alpha}}(u)\right\} \\
&\cdot f_{\boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|u) f_{T - \boldsymbol{\alpha}^T\boldsymbol{X}}(u)\Big\}\Big|_{\boldsymbol{\alpha} = \boldsymbol{\alpha}_0} du\, d\boldsymbol{x}.
\end{aligned}$$

Since the last two terms vanish because the integrands become zero in that case. Note that

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(u) = &\int \boldsymbol{y} f_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{y}) f_{\boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{y}|u)\, d\boldsymbol{y} \\
&+ \int F_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{y}) \frac{\partial}{\partial \boldsymbol{\alpha}} f_{\boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{y}|u)\, d\boldsymbol{y},
\end{aligned}$$

implying that, at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$,

$$\frac{\partial}{\partial \boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(u)\,\Big|_{\boldsymbol{\alpha} = \boldsymbol{\alpha}_0} = f_0(u)\mathbb{E}(\boldsymbol{X}|T - \boldsymbol{\alpha}_0^T\boldsymbol{X} = u).$$

Since

$$\begin{aligned}
&\int_{u = F_0^{-1}(\epsilon)}^{F_0^{-1}(1-\epsilon)} \int \boldsymbol{x}\left\{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|T - \boldsymbol{\alpha}_0^T\boldsymbol{X} = u)\right\}^T f_{\boldsymbol{X}|T - \boldsymbol{\alpha}_0^T\boldsymbol{X}}(\boldsymbol{x}|u)\, f_0(u) f_{T - \boldsymbol{\alpha}_0^T\boldsymbol{X}}(u) d\boldsymbol{x}\, du \\
&= \mathbb{E}_\epsilon\left[\boldsymbol{X}\{\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}|T - \boldsymbol{\alpha}_0^T\boldsymbol{X})\}^T f_0(T - \boldsymbol{\alpha}_0^T\boldsymbol{X})\right],
\end{aligned}$$

Lemma 2.4.1 now follows. □

A.2.1 Proof of existence of a crossing of zero

Proof of Theorem 2.4.1, Part 1 (Existence of a root). Consider the score function

$$\psi_{1\epsilon,n}(\boldsymbol{\alpha}) = \int_{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\big\{\delta - \hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta),$$

where $\hat{F}_{n,\boldsymbol{\alpha}}$ is the nonparametric maximum likelihood estimator (MLE) of the error distribution. According to the discussion in Section 2.4.1 we have to show that there exists a point $\hat{\boldsymbol{\alpha}}_n$ such that

$$\psi_{1\epsilon,n}(\boldsymbol{\alpha}) = \int_{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\big\{\delta - \hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta),$$

has a zero-crossing at $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}_n$. We have:

$$\begin{aligned}
\psi_{1\epsilon,n}(\boldsymbol{\alpha}) &= \int_{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\big\{\delta - F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta)\\
&\quad + \int_{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\big\{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - \hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta)\\
&= \int_{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\big\{\delta - F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta)\\
&\quad + \int_{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\big\{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - \hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\}\, d(\mathbb{P}_n - P_0)(\boldsymbol{x},t,\delta)\\
&\quad + \int_{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\big\{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - \hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\} dP_0(\boldsymbol{x},t,\delta).
\end{aligned} \tag{A.2.2}$$

Let $\mathcal{F}$ be the set of piecewise constant distribution functions with finitely many jumps (like the MLE $\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}$), and let, for $\boldsymbol{\alpha}\in\Theta$, $\mathcal{K}$ be the set of functions

$$\mathcal{K} = \big\{(\boldsymbol{x},t,\delta) \mapsto \boldsymbol{x}\big\{\delta - F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\} 1_{[\epsilon,1-\epsilon]}\left(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right) : F\in\mathcal{F}, \boldsymbol{\alpha}\in\Theta\big\}.$$

We add the function

$$(\boldsymbol{x},t,\delta) \mapsto \boldsymbol{x}\big\{\delta - F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\} 1_{[\epsilon,1-\epsilon]}\left(F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right),$$

to $\mathcal{K}$. We denote by $H_B(\zeta,\mathcal{K},L_2(P_0))$ the bracketing ζ-entropy w.r.t. the L_2-distance d, defined by

$$d(k_1,k_2)^2 = \int \|k_1 - k_2\|^2\, dP_0, \qquad k_1,k_2\in\mathcal{K}.$$

Note that

$$\boldsymbol{x}\big\{\delta - F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\} 1_{[\epsilon,1-\epsilon]}\left(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right) = f_{1,\boldsymbol{\alpha}}(\boldsymbol{x},t,\delta) f_{2,\boldsymbol{\alpha}}(\boldsymbol{x},t,\delta),$$

where

$$f_{1,\boldsymbol{\alpha}}(\boldsymbol{x},t,\delta)=\boldsymbol{x}\big\{\delta-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\},$$

and

$$f_{2,\boldsymbol{\alpha}}(\boldsymbol{x},t,\delta)=1_{[\epsilon,1-\epsilon]}\left(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right).$$

Since t and $\boldsymbol{x}$ vary over a bounded region and, by Assumption A4, $F_{\boldsymbol{\alpha}}$ is of bounded variation, $f_{1,\boldsymbol{\alpha}}$ is of bounded variation. Moreover,

$$f_{2,\boldsymbol{\alpha}}(\boldsymbol{x},t,\delta)=1_{[\epsilon,1-\epsilon]}\left(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right)=1_{[\epsilon,1]}\left(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right)-1_{(1-\epsilon,1]}\left(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right).$$

Since F is monotone, we have:

$$\begin{aligned}1_{[\epsilon,1]}&\left(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right)-1_{(1-\epsilon,1]}\left(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right)\\&=1_{[a_{\epsilon,F},M]}(t-\boldsymbol{\alpha}^T\boldsymbol{x})-1_{(b_{\epsilon,F},M]}(t-\boldsymbol{\alpha}^T\boldsymbol{x}),\end{aligned}\tag{A.2.3}$$

for points $a_{\epsilon,F}\le b_{\epsilon,F}$, where M is an upper bound for the values of $t-\boldsymbol{\alpha}^T\boldsymbol{x}$. Hence $f_{2,\boldsymbol{\alpha}}$ is also a function of uniformly bounded variation.
We therefore get, using Lemma A.1.1, that

$$\sup_{\zeta>0}\zeta H_B\left(\zeta,\mathcal{K},L_2(P_0)\right)=O(1),$$

which implies that

$$\int_0^{\zeta}H_B\left(u,\mathcal{K},L_2(P_0)\right)^{1/2}du=O\left(\zeta^{1/2}\right),\qquad \zeta>0.$$

This implies:

$$\begin{aligned}&\int_{\hat F_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]}\boldsymbol{x}\big\{\delta-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\}\,d\mathbb{P}_n(\boldsymbol{x},t,\delta)\\&=\int_{\hat F_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]}\boldsymbol{x}\big\{\delta-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\}\,dP_0(\boldsymbol{x},t,\delta)\\&\qquad+\int_{\hat F_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]}\boldsymbol{x}\big\{\delta-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\}\,d\big(\mathbb{P}_n-P_0\big)(\boldsymbol{x},t,\delta)\\&=\int_{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]}\boldsymbol{x}\big\{\delta-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\}\,dP_0(\boldsymbol{x},t,\delta)\\&\qquad+\int_{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]}\boldsymbol{x}\big\{\delta-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\}\,d\big(\mathbb{P}_n-P_0\big)(\boldsymbol{x},t,\delta)+o_p(1)\\&=\psi_{1\epsilon}(\boldsymbol{\alpha})+o_p(1),\end{aligned}$$

uniformly in $\boldsymbol{\alpha} \in \Theta$, by the convergence in probability (and almost surely) of $\hat{F}_{n,\boldsymbol{\alpha}}$ to $F_{\boldsymbol{\alpha}}$, where we use Lemma A.0.1 for the second term on the right-hand side of the first equality to make the transition of the integration region $\hat{F}_{n,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \in [\epsilon, 1-\epsilon]$ to $F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \in [\epsilon, 1-\epsilon]$.

For the second term of (A.2.2) we argue similarly, this time using the function class

$$\mathcal{K}' = \big\{(\boldsymbol{x}, t, \delta) \mapsto \boldsymbol{x}\big\{F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) - F(t - \boldsymbol{\alpha}^T\boldsymbol{x})\big\}1_{[\epsilon,1-\epsilon]}\left(F(t - \boldsymbol{\alpha}^T\boldsymbol{x})\right) : \\ F \in \mathcal{F},\, \boldsymbol{\alpha} \in \Theta\big\},$$

to which we add the function that is identically zero. This implies that these terms are $o_p(1)$. For the third term of (A.2.2) we get by an application of the Cauchy-Schwarz inequality that, uniformly in $\boldsymbol{\alpha}$,

$$\begin{aligned}
&\int_{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\big\{F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) - \hat{F}_{n,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\big\} dP_0(\boldsymbol{x}, t, \delta) \\
&\le \Bigg(\int_{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \|\boldsymbol{x}\|^2 dP_0(\boldsymbol{x}, t, \delta) \\
&\qquad \cdot \int_{\hat{F}_{n,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \big\{F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) - \hat{F}_{n,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\big\}^2 dP_0(\boldsymbol{x}, t, \delta)\Bigg)^{1/2} \\
&= O_p(n^{-1/3}).
\end{aligned}$$

The conclusion is that

$$\psi_{1\epsilon,n}(\boldsymbol{\alpha}) = \psi_{1\epsilon}(\boldsymbol{\alpha}) + o_p(1),$$

uniformly in $\boldsymbol{\alpha} \in \Theta$.

[Existence of $\hat{\boldsymbol{\alpha}}_n$:] Let $\psi_{1\epsilon}$ be the population version of the statistic $\psi_{1\epsilon,n}$ defined by (A.2.1). We have:

$$\psi_{1\epsilon}(\boldsymbol{\alpha}_0) = \mathbf{0}.$$

Furthermore,

$$\psi_{1\epsilon,n}(\boldsymbol{\alpha}) = \psi'_{1\epsilon}(\boldsymbol{\alpha}_0)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + R_n(\boldsymbol{\alpha}), \tag{A.2.4}$$

where $R_n(\boldsymbol{\alpha}) = o_p(1) + o(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)$, and where the $o_p(1)$ term is uniform in $\boldsymbol{\alpha} \in \Theta$. Note that $\psi'_{1\epsilon}(\boldsymbol{\alpha}_0)$ is by definition nonsingular.

We now define, for $h > 0$, the functions

$$\tilde{R}_{n,h}(\boldsymbol{\alpha}) = h^{-d} \int K_h(u_1 - \alpha_1) \dots K_h(u_d - \alpha_d)\, R_n(u_1, \dots, u_d)\, du_1 \dots du_d,$$

where d is the dimension of Θ and

$$K_h(x) = h^{-1}K(x/h), \qquad x \in \mathbb{R},$$

letting K be one of the usual smooth kernels with support $[-1, 1]$, like the Triweight kernel that we used in the simulations.
Furthermore, we define:

$$\tilde{\psi}_{1\epsilon,n,h}(\boldsymbol{\alpha}) = \psi'_{1\epsilon}(\boldsymbol{\alpha}_0)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + \tilde{R}_{nh}(\boldsymbol{\alpha}).$$

Clearly:

$$\lim_{h \downarrow 0} \tilde{\psi}_{1\epsilon,n,h}(\boldsymbol{\alpha})) = \psi_{1\epsilon,n}(\boldsymbol{\alpha}) \qquad \text{and} \qquad \lim_{h \downarrow 0} \tilde{R}_{nh}(\boldsymbol{\alpha}) = R_n(\boldsymbol{\alpha}),$$

for each continuity point $\boldsymbol{\alpha}$ of $\psi_{1\epsilon,n}$.
We now reparametrize, defining

$$\boldsymbol{\gamma} = \psi'_{1\epsilon}(\boldsymbol{\alpha}_0)\boldsymbol{\alpha} \qquad \text{and} \qquad \boldsymbol{\gamma}_0 = \psi'_{1\epsilon}(\boldsymbol{\alpha}_0)\boldsymbol{\alpha}_0.$$

This gives:

$$\psi'_{1\epsilon}(\boldsymbol{\alpha}_0)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + \tilde{R}_{nh}(\boldsymbol{\alpha}) = \boldsymbol{\gamma} - \boldsymbol{\gamma}_0 + \tilde{R}_{nh}\left(\psi'_{1\epsilon}(\boldsymbol{\alpha}_0)^{-1}\boldsymbol{\gamma}\right).$$

By (A.2.4), the mapping

$$\boldsymbol{\gamma} \mapsto \boldsymbol{\gamma}_0 - R_n\left(\psi'_{1\epsilon}(\boldsymbol{\alpha}_0)^{-1}\boldsymbol{\gamma}\right),$$

maps, for each $\eta > 0$, the ball $B_\eta(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \le \eta\}$ into $B_{\eta/2}(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \le \eta/2\}$ for all large n, with probability tending to one, where $\|\cdot\|$ denotes the Euclidean norm, implying that the *continuous* map

$$\boldsymbol{\gamma} \mapsto \boldsymbol{\gamma}_0 - \tilde{R}_{nh}\left(\psi'_{1\epsilon}(\boldsymbol{\alpha}_0)^{-1}\boldsymbol{\gamma}\right),$$

maps $B_\eta(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \le \eta\}$ into itself for all large n and small h. So for large n and small h there is, by Brouwer's fixed point theorem a point $\boldsymbol{\gamma}_{nh}$ such that

$$\boldsymbol{\gamma}_{nh} = \boldsymbol{\gamma}_0 - \tilde{R}_{nh}\left(\psi'_{1\epsilon}(\boldsymbol{\alpha}_0)^{-1}\boldsymbol{\gamma}_{nh}\right).$$

Defining $\boldsymbol{\alpha}_{nh} = \psi'_{1\epsilon}(\boldsymbol{\alpha}_0)^{-1}\boldsymbol{\gamma}_{nh}$, we get:

$$\tilde{\psi}_{1\epsilon,n,h}(\boldsymbol{\alpha}_{nh}) = \psi'_{1\epsilon}(\boldsymbol{\alpha}_0)(\boldsymbol{\alpha}_{nh} - \boldsymbol{\alpha}_0) + \tilde{R}_{nh}(\boldsymbol{\alpha}_{nh}) = \mathbf{0}. \tag{A.2.5}$$

By compactness, $(\boldsymbol{\alpha}_{n,1/k})_{k=1}^{\infty}$ must have a subsequence $(\boldsymbol{\alpha}_{n,1/k_i})$ with a limit $\tilde{\boldsymbol{\alpha}}_n$, as $i \to \infty$. We show that each component of $\psi_{1\epsilon,n}$ has a crossing of zero at $\tilde{\boldsymbol{\alpha}}_n$.

Suppose that the jth component $\psi_{1\epsilon,n,j}$ of $\psi_{1\epsilon,n}$ does not have a crossing of zero at $\tilde{\boldsymbol{\alpha}}_n$. Then there must be an open ball $B_\delta(\tilde{\boldsymbol{\alpha}}_n) = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}}_n\| < \delta\}$ of $\tilde{\boldsymbol{\alpha}}_n$ such that $\psi_{1\epsilon,n,j}$ has a constant sign in $B_\delta(\tilde{\boldsymbol{\alpha}}_n)$, say $\psi_{1\epsilon,n,j}(\boldsymbol{\alpha}) > 0$ for $\boldsymbol{\alpha} \in B_\delta(\tilde{\boldsymbol{\alpha}}_n)$. Since $\psi_{1\epsilon,n,j}$ only has finitely many values, this means that

$$\psi_{1\epsilon,n,j}(\boldsymbol{\alpha}) \ge c > 0, \qquad \text{for all } \boldsymbol{\alpha} \in B_\delta(\tilde{\boldsymbol{\alpha}}_n),$$

for some $c > 0$. This means that the jth component $\tilde{\psi}_{1\epsilon,n,h,j}$ of $\tilde{\psi}_{1\epsilon,n,h}$ satisfies

$$\begin{aligned}
\tilde{\psi}_{1\epsilon,n,h,j}(\boldsymbol{\alpha}) &= \left[\psi'_{1\epsilon}(\boldsymbol{\alpha}_0)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)\right]_j + \tilde{R}_{nh,j}(\boldsymbol{\alpha}) \\
&= h^{-d} \int \Big\{\left[\psi'_{1\epsilon}(\boldsymbol{\alpha}_0)(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)\right]_j + R_{nj}(u_1, \dots, u_d)\Big\} \\
&\qquad\qquad \cdot K_h(u_1 - \alpha_1) \dots K_h(u_d - \alpha_d)\, du_1 \dots du_d \\
&\ge h^{-d} \int \Big\{\left[\psi'_{1\epsilon}(\boldsymbol{\alpha}_0)(u - \boldsymbol{\alpha}_0)\right]_j + R_{nj}(u_1, \dots, u_d)\Big\} \\
&\qquad\qquad \cdot K_h(u_1 - \alpha_1) \dots K_h(u_d - \alpha_d)\, du_1 \dots du_d - c/2 \\
&\ge c\, h^{-d} \int K_h(u_1 - \alpha_1) \dots K_h(u_d - \alpha_d)\, du_1 \dots du_d - c/2 \\
&= c/2,
\end{aligned}$$

for $\boldsymbol{\alpha} \in B_{\delta/2}(\tilde{\boldsymbol{\alpha}}_n)$ and sufficiently small h, contradicting (A.2.5), since $\boldsymbol{\alpha}_{nh}$, for $h = 1/k_i$, belongs to $B_{\delta/2}(\tilde{\boldsymbol{\alpha}}_n)$ for large k_i. □

A.2.2 Proof of consistency of the SSE

Proof of Theorem 2.4.1, Part 2 (Consistency). We assume that $\hat{\boldsymbol{\alpha}}_n$ is contained in the compact set Θ, and hence the sequence $(\hat{\boldsymbol{\alpha}}_n)$ has a subsequence $(\hat{\boldsymbol{\alpha}}_{n_k} = \hat{\boldsymbol{\alpha}}_{n_k}(\omega))$, converging to an element $\boldsymbol{\alpha}_*$. If $\hat{\boldsymbol{\alpha}}_{n_k} = \hat{\boldsymbol{\alpha}}_{n_k}(\omega) \longrightarrow \boldsymbol{\alpha}_*$, we get by Lemma 2.3.1,

$$\hat{F}_{n_k,\hat{\boldsymbol{\alpha}}_{n_k}}(t - \hat{\boldsymbol{\alpha}}_{n_k}^T \boldsymbol{x}) \longrightarrow F_{\boldsymbol{\alpha}_*}(t - \boldsymbol{\alpha}_*^T \boldsymbol{x}),$$

where $F_{\boldsymbol{\alpha}}$ is defined in (2.3.1). In the limit we get therefore the relation:

$$\begin{aligned}
&\lim_{k\to\infty} \int_{\hat{F}_{n_k,\hat{\boldsymbol{\alpha}}_{n_k}}(t - \hat{\boldsymbol{\alpha}}_{n_k}^T \boldsymbol{x}) \in [\epsilon, 1-\epsilon]} \boldsymbol{x}\big\{\delta - F_{n_k,\hat{\boldsymbol{\alpha}}_{n_k}}(t - \hat{\boldsymbol{\alpha}}_{n_k}^T \boldsymbol{x})\big\} d\mathbb{P}_{n_k}(\boldsymbol{x}, t, \delta) \\
&= \int_{F_{\boldsymbol{\alpha}_*}(t - \boldsymbol{\alpha}_*^T \boldsymbol{x}) \in [\epsilon, 1-\epsilon]} \boldsymbol{x}\big\{F_0(t - \boldsymbol{\alpha}_0^T \boldsymbol{x}) - F_{\boldsymbol{\alpha}_*}(t - \boldsymbol{\alpha}_*^T \boldsymbol{x})\big\}\, dG(\boldsymbol{x}, t) = \boldsymbol{0},
\end{aligned}$$

using that, in the limit, the crossing of zero becomes a root of the continuous limiting function. Consider:

$$\int_{F_{\boldsymbol{\alpha}_*}(t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\big\{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})-F_{\boldsymbol{\alpha}_*}(t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\big\}\,dG(\boldsymbol{x},t)$$
$$=\int_{F_{\boldsymbol{\alpha}_*}(t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\big\{F_0(t-\boldsymbol{\alpha}_*^T\boldsymbol{x}+(\boldsymbol{\alpha}_*-\boldsymbol{\alpha}_0)^T\boldsymbol{x})-F_{\boldsymbol{\alpha}_*}(t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\big\}\,dG(\boldsymbol{x},t).$$

Since

$$F_{\boldsymbol{\alpha}_*}(t-\boldsymbol{\alpha}_*^T\boldsymbol{x})$$
$$=\int F_0(t-\boldsymbol{\alpha}_*^T\boldsymbol{x}+(\boldsymbol{\alpha}_*-\boldsymbol{\alpha}_0)^T\boldsymbol{y})f_{\boldsymbol{X}|T-\boldsymbol{\alpha}_*^T\boldsymbol{X}}(\boldsymbol{y}|T-\boldsymbol{\alpha}_*^T\boldsymbol{X}=t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\,d\boldsymbol{y},$$

we get:

$$(\boldsymbol{\alpha}_*-\boldsymbol{\alpha}_0)^T\int_{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\Big\{F_0(t-\boldsymbol{\alpha}_*^T\boldsymbol{x}+(\boldsymbol{\alpha}_*-\boldsymbol{\alpha}_0)^T\boldsymbol{x})-F_{\boldsymbol{\alpha}_*}(t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\Big\}$$
$$\cdot\, dG(\boldsymbol{x},t)$$
$$=\int_{F_{\boldsymbol{\alpha}_*}(t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\in[\epsilon,1-\epsilon]}(\boldsymbol{\alpha}_*-\boldsymbol{\alpha}_0)^T\boldsymbol{x}\Big\{F_0(t-\boldsymbol{\alpha}_*^T\boldsymbol{x}+(\boldsymbol{\alpha}_*-\boldsymbol{\alpha}_0)^T\boldsymbol{x})$$
$$-\int F_0(t-\boldsymbol{\alpha}_*^T\boldsymbol{x}+(\boldsymbol{\alpha}_*-\boldsymbol{\alpha}_0)^T\boldsymbol{y})f_{\boldsymbol{X}|T-\boldsymbol{\alpha}_*^T\boldsymbol{X}}(\boldsymbol{y}|T-\boldsymbol{\alpha}_*^T\boldsymbol{X}=t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\,d\boldsymbol{y}\Big\}$$
$$\cdot\, dG(\boldsymbol{x},t)$$
$$=\int_{F_{\boldsymbol{\alpha}_*}(u)\in[\epsilon,1-\epsilon]}\text{Cov}\Big\{(\boldsymbol{\alpha}_*-\boldsymbol{\alpha}_0)^T\boldsymbol{X},F_0(u+(\boldsymbol{\alpha}_*-\boldsymbol{\alpha}_0)^T\boldsymbol{X})\;\Big|\;T-\boldsymbol{\alpha}_*^T\boldsymbol{X}=u\Big\}$$
$$\cdot\, f_{T-\boldsymbol{\alpha}_*^T\boldsymbol{X}}(u)\,du$$
$$=0.$$

We first note that by Lemma 2.4.1 the integrand is positive for all $\boldsymbol{\alpha}_*\in\Theta$. Suppose that $\boldsymbol{\alpha}_*\neq\boldsymbol{\alpha}_0$, then this integral can only be zero if $\text{Cov}((\boldsymbol{\alpha}_*-\boldsymbol{\alpha}_0)^T\boldsymbol{X},F_0(u+(\boldsymbol{\alpha}_*-\boldsymbol{\alpha}_0)^T\boldsymbol{X})|T-\boldsymbol{\alpha}_*^T\boldsymbol{X}=u)$ is zero for all u such that $F_{\boldsymbol{\alpha}_*}(u)\in[\epsilon,1-\epsilon]$, if $f_{T-\boldsymbol{\alpha}_*^T\boldsymbol{X}}(u)$ stays away from zero on this region (Assumptions A3), using continuity of the functions in the integrand (Assumptions A5) and the nonnegativity of the conditional covariance function (see also Remark 2.4.2). Since this is excluded by the condition that the covariance $\text{Cov}(X,F_0(u+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{X})|T-\boldsymbol{\alpha}^T\boldsymbol{X}=u)$ is continuous in u and not identically zero for u in the region $\{u:\epsilon\le F_{\boldsymbol{\alpha}}(u)\le 1-\epsilon\}$, for each $\boldsymbol{\alpha}\in\Theta$, we must have: $\boldsymbol{\alpha}_*=\boldsymbol{\alpha}_0$. □

A.2.3 Proof of asymptotic normality of the SSE

Proof of Theorem 2.4.1, Part 3 (Asymptotic Normality). Before working out the details, we give a kind of "road map" for the proof of Theorem 2.4.1, Part 3.

1. We define $\psi_{1\epsilon,n}$ at $\hat{\boldsymbol{\alpha}}_n$ by putting

$$\psi_{1\epsilon,n}(\hat{\boldsymbol{\alpha}}_n) = \mathbf{0}. \tag{A.2.6}$$

Note that, with this definition, $\psi_{1\epsilon,n}(\hat{\boldsymbol{\alpha}}_n)$ is in dimension 1 just the convex combination of the left and right limit at $\hat{\boldsymbol{\alpha}}_n$:

$$\psi_{1\epsilon,n}(\hat{\boldsymbol{\alpha}}_n) = \gamma\psi_{1\epsilon,n}(\hat{\boldsymbol{\alpha}}_n-) + (1-\gamma)\psi_{1\epsilon,n}(\hat{\boldsymbol{\alpha}}_n+) = 0, \tag{A.2.7}$$

where we can choose $\gamma \in [0,1]$ in such a way that (A.2.7) holds. In dimension d higher than one, we can also define $\psi_{1\epsilon,n}$ at $\hat{\boldsymbol{\alpha}}_n$ by (A.2.6) and use the representation of the components as a convex combination since we have a crossing of zero component wise. Since the following asymptotic representations are also valid for one-sided limits as used in (A.2.7) we can use Definition (A.2.6) and assume $\psi_{1\epsilon,n}(\hat{\boldsymbol{\alpha}}_n) = \mathbf{0}$.

We show:

$$\begin{aligned}\psi_{1\epsilon,n}(\hat{\boldsymbol{\alpha}}_n) = &\int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{\boldsymbol{x} - \phi_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\Big\} \\ &\qquad\qquad \cdot \Big\{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\Big\}\, dP_0(\boldsymbol{x},t,\delta) \\ &+ \int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{\boldsymbol{x} - \phi_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\Big\}\Big\{\delta - F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\Big\} \\ &\qquad\qquad\qquad d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x},t,\delta) \\ &+ o_p\Big(n^{-1/2} + \hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\Big),\end{aligned} \tag{A.2.8}$$

where

$$\phi_0(u) = \phi_{\boldsymbol{\alpha}_0}(u),$$

and where $\phi_{\boldsymbol{\alpha}}$ is defined by

$$\phi_{\boldsymbol{\alpha}}(u) = \mathbb{E}\left\{\boldsymbol{X}|T - \boldsymbol{\alpha}^T\boldsymbol{X} = u\right\}. \tag{A.2.9}$$

Since $\hat{\boldsymbol{\alpha}}_n \xrightarrow{p} \boldsymbol{\alpha}_0$ and

$$\begin{aligned}&\int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{\boldsymbol{x} - \phi_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\Big\} \\ &\qquad\qquad \cdot \Big\{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\Big\}\, dP_0(\boldsymbol{x},t,\delta) \\ &= \psi_{1\epsilon}'(\boldsymbol{\alpha}_0)\left(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\right) + o_p\left(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\right),\end{aligned}$$

this yields, using the invertibility of $\psi'_{1\epsilon}(\boldsymbol{\alpha}_0)$,

$$\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) \\
&= -\psi'_{1\epsilon}(\boldsymbol{\alpha}_0)^{-1}\left\{\sqrt{n}\int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]}\left\{\boldsymbol{x} - \phi_{\boldsymbol{\alpha}_0}(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right\}\right. \\
&\qquad\qquad\qquad\left.\cdot\left\{\delta - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right\} d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x}, t, \delta)\right\} \\
&\qquad + o_p\left(1 + \sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)\right).
\end{aligned}$$

As a consequence, the result of Theorem 2.4.1 follows, since

$$\begin{aligned}
&\sqrt{n}\int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]}\left\{\boldsymbol{x} - \phi_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right\}\left\{\delta - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right\} d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x}, t, \delta) \\
&\stackrel{d}{\to} N(\boldsymbol{0}, \boldsymbol{B}).
\end{aligned}$$

2. To show that (A.2.8) holds, we need entropy results for the functions $u \mapsto \hat{F}_{n,\boldsymbol{\alpha}}(u)$ and $u \mapsto \bar{\phi}_{\boldsymbol{\alpha},\hat{F}_{n,\boldsymbol{\alpha}}}(u)$ (see (A.2.10) below). We also have to deal with the simpler parametric functions $F_{\boldsymbol{\alpha}}$ and $\phi_{\boldsymbol{\alpha}}$, parametrized by the finite dimensional parameter $\boldsymbol{\alpha}$, which are the population equivalents of $\hat{F}_{n,\boldsymbol{\alpha}}$ and $\bar{\phi}_{\boldsymbol{\alpha},\hat{F}_{n,\boldsymbol{\alpha}}}$.

3. The result will then follow from the properties of $F_{\boldsymbol{\alpha}}$ and $\phi_{\boldsymbol{\alpha}}$, together with the closeness of $\hat{F}_{n,\boldsymbol{\alpha}}$ to $F_{\boldsymbol{\alpha}}$ and $\bar{\phi}_{\boldsymbol{\alpha},\hat{F}_{n,\boldsymbol{\alpha}}}$ to $\phi_{\boldsymbol{\alpha}}$, respectively, and the convergence of $\hat{\boldsymbol{\alpha}}_n$ to $\boldsymbol{\alpha}_0$.

Let $\bar{\phi}_{\hat{\boldsymbol{\alpha}}_n,\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}}$ be a (random) piecewise constant version of $\phi_{\hat{\boldsymbol{\alpha}}_n}$, where, for a piecewise constant distribution function F with finitely many jumps at $\tau_1 < \tau_2 < \dots$, the function $\bar{\phi}_{\boldsymbol{\alpha},F}$ is defined in the following way.

$$\bar{\phi}_{\boldsymbol{\alpha},F}(u) = \begin{cases} \phi_{\boldsymbol{\alpha}}(\tau_i), & \text{if } F_{\boldsymbol{\alpha}}(u) > F(\tau_i),\ u \in [\tau_i, \tau_{i+1}), \\ \phi_{\boldsymbol{\alpha}}(s), & \text{if } F_{\boldsymbol{\alpha}}(u) = F(s),\ \text{for some } s \in [\tau_i, \tau_{i+1}), \\ \phi_{\boldsymbol{\alpha}}(\tau_{i+1}), & \text{if } F_{\boldsymbol{\alpha}}(u) < F(\tau_i),\ u \in [\tau_i, \tau_{i+1}). \end{cases} \tag{A.2.10}$$

We can write:

$$\begin{aligned}
\psi_{1\epsilon,n}(\hat{\boldsymbol{\alpha}}_n) &= \int_{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x}\big\{\delta - \hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\big\}\,d\mathbb{P}_n(\boldsymbol{x},t,\delta)\\
&= \int_{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \big\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\big\}\big\{\delta - \hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\big\}\\
&\qquad\qquad\cdot\, d\mathbb{P}_n(\boldsymbol{x},t,\delta)\\
&\quad + \int_{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \big\{\phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \bar\phi_{\hat{\boldsymbol{\alpha}}_n,\hat F_{n,\hat{\boldsymbol{\alpha}}_n}}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\big\}\\
&\qquad\qquad\cdot\big\{\delta - \hat F_{n,\boldsymbol{\alpha}}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\big\}\,d\mathbb{P}_n(\boldsymbol{x},t,\delta)\\
&= I + II,
\end{aligned}$$

using

$$\begin{aligned}
&\int_{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \bar\phi_{\hat{\boldsymbol{\alpha}}_n,\hat F_{n,\hat{\boldsymbol{\alpha}}_n}}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\big\{\delta - \hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\big\}\,d\mathbb{P}_n(\boldsymbol{x},t,\delta)\\
&= \boldsymbol{0},
\end{aligned}$$

by the definition of the MLE $\hat F_{n,\hat{\boldsymbol{\alpha}}_n}$ as the slope of the greatest convex minorant of the corresponding cusum diagram, based on the values of the Δ_i in the ordering of the $T_i - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}_i$ (see also Lemma A.5 on p.380 of Groeneboom et al. (2010)).
We first show that

$$II = o_p\left(n^{-1/2} + (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)\right).$$

Since the function $u \mapsto \phi_{\boldsymbol{\alpha}}(u)$ has a totally bounded derivative (as a consequence of (A.2.9) and Assumption A5), we can bound the Euclidean norm of the differences $\phi_{\boldsymbol{\alpha}}(u) - \bar\phi_{\boldsymbol{\alpha},\hat F_{n,\boldsymbol{\alpha}}}(u)$ above by a constant times $|\hat F_{n,\boldsymbol{\alpha}}(u) - F_{\boldsymbol{\alpha}}(u)|$, for $u \in A_{\epsilon,\boldsymbol{\alpha}}$ (see Assumption A2), i.e.,

$$\|\phi_{\boldsymbol{\alpha}}(u) - \bar\phi_{\boldsymbol{\alpha},\hat F_{n,\boldsymbol{\alpha}}}(u)\| \le K_{\boldsymbol{\alpha}}|\hat F_{n,\boldsymbol{\alpha}}(u) - F_{\boldsymbol{\alpha}}(u)|,$$

for some constant $K_{\boldsymbol{\alpha}} > 0$ where the constant $K_{\boldsymbol{\alpha}}$ depends on $\boldsymbol{\alpha}$ through $f_{\boldsymbol{\alpha}}$ (see for this technique for example (10.64) in Groeneboom and Jongbloed (2014)). By Assumption A2 we know that $f_{\boldsymbol{\alpha}}$ is continuous for all $\boldsymbol{\alpha}_n \in \Theta$ such that we can find a constant $K > 0$ not depending on $\boldsymbol{\alpha}$, satisfying,

$$\|\phi_{\boldsymbol{\alpha}}(u) - \bar\phi_{\boldsymbol{\alpha},\hat F_{n,\boldsymbol{\alpha}}}(u)\| \le K|\hat F_{n,\boldsymbol{\alpha}}(u) - F_{\boldsymbol{\alpha}}(u)|, \tag{A.2.11}$$

uniformly in $\boldsymbol{\alpha}_n \in \Theta$. Note that we also need $f_{\boldsymbol{\alpha}}(u) > 0$ for applying this, which is ensured by Assumption A2. We have:

$$
\begin{aligned}
II &= \int_{\hat F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\in[\epsilon,1-\epsilon]} \Big\{\phi_{\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)-\bar\phi_{\hat{\boldsymbol\alpha}_n,\hat F_{n,\hat{\boldsymbol\alpha}_n}}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\Big\}\\
&\qquad\qquad\cdot\Big\{\delta-\hat F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\Big\}\,d\mathbb P_n(\boldsymbol x,t,\delta)\\
&= \int_{\hat F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\in[\epsilon,1-\epsilon]} \Big\{\phi_{\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)-\bar\phi_{\hat{\boldsymbol\alpha}_n,\hat F_{n,\hat{\boldsymbol\alpha}_n}}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\Big\}\\
&\qquad\qquad\cdot\Big\{\delta-\hat F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\Big\}\,d(\mathbb P_n-P_0)(\boldsymbol x,t,\delta)\\
&\quad+\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n}(u)\in[\epsilon,1-\epsilon]}\Big\{\phi_{\hat{\boldsymbol\alpha}_n}(u)-\bar\phi_{\hat{\boldsymbol\alpha}_n,\hat F_{n,\hat{\boldsymbol\alpha}_n}}(u)\Big\}\Big\{F_{\hat{\boldsymbol\alpha}_n}(u)-\hat F_{n,\hat{\boldsymbol\alpha}_n}(u)\Big\}\\
&\qquad\qquad\cdot f_{T-\hat{\boldsymbol\alpha}_n^T\boldsymbol X}(u)\,du\\
&\quad+\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\in[\epsilon,1-\epsilon]} \Big\{\phi_{\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)-\bar\phi_{\hat{\boldsymbol\alpha}_n,\hat F_{n,\hat{\boldsymbol\alpha}_n}}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\Big\}\\
&\qquad\qquad\cdot\Big\{F_0(t-\boldsymbol\alpha_0^T\boldsymbol x)-F_{\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\Big\}\,dP_0(\boldsymbol x,t,\delta)\\
&= II_a+II_b+II_c.
\end{aligned}
$$

First consider II_a. Let $\mathcal F$ be the set of piecewise constant distribution functions with finitely many jumps (like the MLE $\hat F_{n,\hat{\boldsymbol\alpha}_n}$), and let $\mathcal K_1$ be the set of functions

$$
\begin{aligned}
\mathcal K_1 = \big\{(\boldsymbol x,t,\delta)\mapsto \big(\phi_{\boldsymbol\alpha}(t-\boldsymbol\alpha^T\boldsymbol x)-\bar\phi_{\boldsymbol\alpha,F}(t-\boldsymbol\alpha^T\boldsymbol x)\big)\big(\delta-F(t-\boldsymbol\alpha^T\boldsymbol x)\big)\\
\cdot 1_{[\epsilon,1-\epsilon]}(F(t-\boldsymbol\alpha^T\boldsymbol x)) : F\in\mathcal F,\ \boldsymbol\alpha\in\Theta\big\},
\end{aligned}
$$

where $\bar\phi_{\boldsymbol\alpha,F}$ is again defined by (A.2.10). We add the function which is identically zero to $\mathcal K_1$.

The functions $u\mapsto F(u)$, for $F\in\mathcal F$ and (as argued above) $u\mapsto\bar\phi_{\boldsymbol\alpha,F}(u)$ are bounded functions of uniformly bounded variation. Note that, for $F_1,F_2\in\mathcal F$,

$$
\begin{aligned}
F_1(t-\boldsymbol\alpha_1^T\boldsymbol x)-F_2(t-\boldsymbol\alpha_2^T\boldsymbol x) &= F_1(t-\boldsymbol\alpha_1^T\boldsymbol x)-F_{\boldsymbol\alpha_1}(t-\boldsymbol\alpha_1^T\boldsymbol x)+F_{\boldsymbol\alpha_1}(t-\boldsymbol\alpha_1^T\boldsymbol x)\\
&\quad-F_{\boldsymbol\alpha_2}(t-\boldsymbol\alpha_2^T\boldsymbol x)+F_{\boldsymbol\alpha_2}(t-\boldsymbol\alpha_2^T\boldsymbol x)-F_2(t-\boldsymbol\alpha_2^T\boldsymbol x),
\end{aligned}
$$

and that (see (2.3.1)):

$$
\begin{aligned}
&\big|F_{\boldsymbol\alpha_1}(t-\boldsymbol\alpha_1^T\boldsymbol x)-F_{\boldsymbol\alpha_2}(t-\boldsymbol\alpha_2^T\boldsymbol x)\big|\\
&=\bigg|\int F_0(t-\boldsymbol\alpha_0^T\boldsymbol x+(\boldsymbol\alpha_1-\boldsymbol\alpha_0)^T(\boldsymbol y-\boldsymbol x))f_{\boldsymbol X|T-\boldsymbol\alpha_1^T\boldsymbol X}(\boldsymbol y|t-\boldsymbol\alpha_1^T\boldsymbol x)\,d\boldsymbol y\\
&\qquad-\int F_0(t-\boldsymbol\alpha_0^T\boldsymbol x+(\boldsymbol\alpha_2-\boldsymbol\alpha_0)^T(\boldsymbol y-\boldsymbol x))f_{\boldsymbol X|T-\boldsymbol\alpha_2^T\boldsymbol X}(\boldsymbol y|t-\boldsymbol\alpha_2^T\boldsymbol x)\,d\boldsymbol y\bigg|\\
&=O\left(|\boldsymbol\alpha_1-\boldsymbol\alpha_2|\right),
\end{aligned}
$$

by Assumption A2 and Assumption A5.
For the indicator function $1_{[\epsilon,1]}\left(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right)$ we get, as in (A.2.3), using the monotonicity of F,

$$\begin{aligned}1_{[\epsilon,1]}\left(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right) &= 1_{[\epsilon,1]}\left(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right) - 1_{(1-\epsilon,1]}\left(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right)\\ &= 1_{[a_{\epsilon,F},M]}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - 1_{(b_{\epsilon,F},M]}(t-\boldsymbol{\alpha}^T\boldsymbol{x}),\end{aligned}$$

for points $a_{\epsilon,F}\le b_{\epsilon,F}$, where M is an upper bound for the values of $t-\boldsymbol{\alpha}^T\boldsymbol{x}$, implying that the function

$$(\boldsymbol{x},t)\mapsto 1_{[\epsilon,1]}\left(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right),$$

is of uniformly bounded variation. So the functions in $\mathcal{K}_1$ are products of functions of uniformly bounded variation, and we therefore get, using Lemma A.1.1

$$\sup_{\zeta>0}\zeta H_B\left(\zeta,\mathcal{K}_1,L_2(P_0)\right)=O(1),$$

which implies:

$$\int_0^\zeta H_B\left(u,\mathcal{K}_1,L_2(P_0)\right)^{1/2}\,du = O\left(\zeta^{1/2}\right),\qquad \zeta>0.$$

Defining:

$$k_{\boldsymbol{\alpha},F}(\boldsymbol{x},t,\delta)=\left(\phi_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})-\bar\phi_{\boldsymbol{\alpha},F}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right)\left(\delta-F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right)1_{[\epsilon,1-\epsilon]}(F(t-\boldsymbol{\alpha}^T\boldsymbol{x})),$$

for $F\in\mathcal{F}$, we get, using (A.2.11),

$$\begin{aligned}&\left\{\int\left\|k_{\hat{\boldsymbol{\alpha}}_n,\hat F_{n,\hat{\boldsymbol{\alpha}}_n}}(\boldsymbol{x},t,\delta)\right\|^2 dP_0(\boldsymbol{x},t,\delta)\right\}^2\\ &\le\int_{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]}\left\|\phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})-\bar\phi_{\hat{\boldsymbol{\alpha}}_n,\hat F_{n,\hat{\boldsymbol{\alpha}}_n}}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\|^2 dP_0(\boldsymbol{x},t,\delta)\\ &\le K\int_{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]}\left\{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})-F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\}^2 dP_0(\boldsymbol{x},t,\delta)\\ &\le K'\int_{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(u)\in[\epsilon,1-\epsilon]}\left\{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(u)-F_{\hat{\boldsymbol{\alpha}}_n}(u)\right\}^2 du\\ &\xrightarrow{p}0,\end{aligned}$$

for constants $K,K'>0$. This implies

$$\sqrt{n}II_a=\sqrt{n}\int k_{\hat{\boldsymbol{\alpha}}_n,\hat F_{n,\hat{\boldsymbol{\alpha}}_n}}(\boldsymbol{x},t,\delta)\,d(\mathbb{P}_n-P_0)(\boldsymbol{x},t,\delta)=o_p(1),\tag{A.2.12}$$

by an application of Lemma A.0.1.

Using (A.2.11), $\|F_{\hat{\boldsymbol{\alpha}}_n} - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}\|_2 = O_p(n^{-1/3})$ and the Cauchy-Schwarz inequality on the second term we get:

$$II_b = O_p(n^{-2/3}).$$

The functions $\phi_{\boldsymbol{\alpha}}$ and $F_{\boldsymbol{\alpha}}$ are of a simple parametric nature, since

$$\phi_{\boldsymbol{\alpha}} = \mathbb{E}(X|T - \boldsymbol{\alpha}^T\boldsymbol{X}),$$

and

$$F_{\boldsymbol{\alpha}}(u) = \int F_0(u + (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^T\boldsymbol{x}) f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{x}|T - \boldsymbol{\alpha}^T\boldsymbol{X} = u)\,d\boldsymbol{x},$$

see (2.3.1). Moreover, since

$$\begin{aligned} F_{\hat{\boldsymbol{\alpha}}_n}(u) &= F_0(u) + (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)^T \int \boldsymbol{x} f_0(u) f_{\boldsymbol{X}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(\boldsymbol{x}|u)\,d\boldsymbol{x} + o_p(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) \\ &= F_0(u) + (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)^T f_0(u)\mathbb{E}\{\boldsymbol{X}|T - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X} = u\} + o_p(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0), \end{aligned}$$

and since the difference $\phi_{\hat{\boldsymbol{\alpha}}_n} - \bar{\phi}_{\hat{\boldsymbol{\alpha}}_n,\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}}$ converges to zero, we get for the third term II_c:

$$\begin{aligned} II_c = &\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\phi_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \bar{\phi}_{\hat{\boldsymbol{\alpha}}_n,\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} \\ &\qquad\cdot \left\{F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x}) - F_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} dP_0(\boldsymbol{x}, t, \delta) \\ = &\; o_p\left(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\right). \end{aligned}$$

We therefore conclude:

$$\psi_{1\epsilon,n}(\hat{\boldsymbol{\alpha}}_n) = I + o_p\left(n^{-1/2} + \hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\right).$$

We now write:

$$\begin{aligned} I = &\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} \\ &\qquad\cdot \left\{\delta - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} d\mathbb{P}_n(\boldsymbol{x}, t, \delta) \\ = &\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} \\ &\qquad\cdot \left\{\delta - F_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} d\mathbb{P}_n(\boldsymbol{x}, t, \delta) \\ &+ \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} \\ &\qquad\cdot \left\{F_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} d\mathbb{P}_n(\boldsymbol{x}, t, \delta) \\ = &\; I_a + I_b. \end{aligned}$$

We next show that I_a is equal to the first two terms on the right-hand side of (A.2.8) and that

$$I_b = o_p\left(n^{-1/2} + (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)\right).$$

We have:

$$\begin{aligned}
I_a &= \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} \\
&\qquad\qquad \cdot\left\{\delta - F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} d\mathbb{P}_n(\boldsymbol{x},t,\delta) \\
&= \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} \\
&\qquad\qquad \cdot\left\{\delta - F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x},t,\delta) \\
&\quad + \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} \\
&\qquad\qquad \cdot\left\{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} dP_0(\boldsymbol{x},t,\delta).
\end{aligned}$$

For the second integral on the right-hand side we get:

$$\begin{aligned}
&\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} \\
&\qquad\qquad \cdot\left\{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} dP_0(\boldsymbol{x},t,\delta) \\
&= \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} \\
&\qquad\qquad \cdot\left\{F_0(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} dP_0(\boldsymbol{x},t,\delta) \\
&\quad + \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} \\
&\qquad\qquad \cdot\left\{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - F_0(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} dP_0(\boldsymbol{x},t,\delta),
\end{aligned}$$

and next we get, using the definition of $\phi_{\boldsymbol{\alpha}}$ given in (A.2.9), for the first integral on the right-hand side of the last display:

$$\begin{aligned}
&\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\}\left\{F_0(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} dP_0(\boldsymbol{x},t,\delta) \\
&= \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)\in[\epsilon,1-\epsilon]} \left\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(u)\right\}\left\{F_0(u) - F_{\hat{\boldsymbol{\alpha}}_n}(u)\right\} f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(u)\, f_{\boldsymbol{X}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(\boldsymbol{x}|u)\, du\, d\boldsymbol{x} \\
&= \mathbf{0}.
\end{aligned}$$

Furthermore, we get by expanding $F_0(t-\boldsymbol{\alpha}^T\boldsymbol{x})$ and by the continuity of $\boldsymbol{\alpha} \mapsto \phi_{\boldsymbol{\alpha}}(u)$ at

$\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$:

$$\int_{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \bigl\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\bigr\}\bigl\{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - F_0(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\bigr\}\,dP_0(\boldsymbol{x},t,\delta)$$
$$= \int_{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \bigl\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\bigr\}(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)^T\boldsymbol{x} f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\,dP_0(\boldsymbol{x},t,\delta)$$
$$\quad + o_p\left(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0\right)$$
$$= \left\{\int_{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \bigl\{\boldsymbol{x} - \phi_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\bigr\}\boldsymbol{x}^T f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\,dP_0(\boldsymbol{x},t,\delta)\right\}(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)$$
$$\quad + o_p\left(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0\right).$$

Finally we get from the consistency of $\hat F_{n,\hat{\boldsymbol{\alpha}}_n}$:

$$\left\{\int_{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \bigl\{\boldsymbol{x} - \phi_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\bigr\}\boldsymbol{x}^T f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\,dP_0(\boldsymbol{x},t,\delta)\right\}(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)$$
$$= \left\{\int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \bigl\{\boldsymbol{x} - \phi_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\bigr\}\boldsymbol{x}^T f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\,dP_0(\boldsymbol{x},t,\delta)\right\}(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)$$
$$\quad + o_p\left(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0\right)$$
$$= \psi_{1\epsilon}'(\boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0) + o_p\left(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0\right).$$

We therefore obtain:

$$I_a$$
$$= \int_{\hat F_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \bigl\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\bigr\}\bigl\{\delta - F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\bigr\}\,d(\mathbb{P}_n-P_0)(\boldsymbol{x},t,\delta)$$
$$\quad + \psi_{1\epsilon}'(\boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0) + o_p\left(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0\right).$$

We now proceed again as before, and define $\mathcal{K}_1'$ to be the set of functions

$$\mathcal{K}_1' = \bigl\{(\boldsymbol{x},t,\delta) \mapsto \bigl(\boldsymbol{x} - \phi_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\bigr)\bigl(\delta - F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\bigr)1_{[\epsilon,1-\epsilon]}(F(t-\boldsymbol{\alpha}^T\boldsymbol{x}))$$
$$: F\in\mathcal{F},\, \boldsymbol{\alpha}\in\Theta\bigr\}.$$

We add the function

$$(\boldsymbol{x},t,\delta) \mapsto \bigl(\boldsymbol{x} - \phi_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\bigr)\bigl(\delta - F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\bigr)1_{[\epsilon,1-\epsilon]}(F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})),$$

to the set $\mathcal{K}_1'$. We therefore get, similarly as before, using Lemma A.1.1, that

$$\sup_{\zeta>0}\zeta H_B\left(\zeta,\mathcal{K}_1',L_2(P_0)\right) = O(1),$$

which implies that

$$\int_0^\zeta H_B\left(u, \mathcal{K}_1', L_2(P_0)\right)^{1/2} du = O\left(\zeta^{1/2}\right), \qquad \zeta > 0.$$

Moreover, we get that

$$\begin{aligned}
&\left(\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right)\left(\delta - F_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right)1_{[\epsilon,1-\epsilon]}\left(\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right)\\
&\qquad - \left(\boldsymbol{x} - \phi_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right)\left(\delta - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right)1_{[\epsilon,1-\epsilon]}\left(F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right)\\
&= \Big\{\left(\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right)\left(\delta - F_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right)\\
&\qquad - \left(\boldsymbol{x} - \phi_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right)\left(\delta - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right)\Big\}1_{[\epsilon,1-\epsilon]}\left(\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right)\\
&\qquad + \left(\boldsymbol{x} - \phi_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right)\left(\delta - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right)\\
&\qquad\qquad \cdot \Big\{1_{[\epsilon,1-\epsilon]}\left(F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right) - 1_{[\epsilon,1-\epsilon]}\left(\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right)\Big\}\\
&= A_n(\boldsymbol{x}, t, \delta) + B_n(\boldsymbol{x}, t, \delta),
\end{aligned}$$

implying that

$$\begin{aligned}
&\int\Big\{\left(\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right)\left(\delta - F_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right)1_{[\epsilon,1-\epsilon]}\left(\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right)\\
&- \left(\boldsymbol{x} - \phi_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right)\left(\delta - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right)1_{[\epsilon,1-\epsilon]}\left(F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\right)\Big\}^2 dP_0(\boldsymbol{x}, t, \delta)\\
&\le 2\int\left\{A_n(\boldsymbol{x}, t, \delta)^2 + B_n(\boldsymbol{x}, t, \delta)^2\right\} dP_0(\boldsymbol{x}, t, \delta) = o_p(1),
\end{aligned}$$

since the integrals w.r.t. A_n^2 and B_n^2 tends to zero using the consistency of $\hat{\boldsymbol{\alpha}}_n$ and $\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}$. Hence we get from Lemma A.0.1:

$$\begin{aligned}
I_a = &\int_{F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x}) \in [\epsilon,1-\epsilon]}\Big\{\boldsymbol{x} - \phi_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\Big\}\Big\{\delta - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\Big\}\, d(\mathbb{P}_n - P_0)(\boldsymbol{x}, t, \delta)\\
&+ \psi_{1\epsilon}'(\boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) + o_p\left(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\right) + o_p\left(n^{-1/2}\right).
\end{aligned}$$

This means that we get the conclusion:

$$\begin{aligned}
&\int_{F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x}) \in [\epsilon,1-\epsilon]}\Big\{\boldsymbol{x} - \phi_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\Big\}\Big\{\delta - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\Big\}\, d\left(\mathbb{P}_n - P_0\right)(\boldsymbol{x}, t, \delta)\\
&= -\psi_{1\epsilon}'(\boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) + o_p\left(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\right) + o_p\left(n^{-1/2}\right),
\end{aligned} \tag{A.2.13}$$

if we can show that I_b is negligible.
Since, by definition of $\phi_{\boldsymbol{\alpha}}$ given in (A.2.9),

$$\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) \in [\epsilon,1-\epsilon]}\Big\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\Big\}\, f_{\boldsymbol{X}|T - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(\boldsymbol{x}|t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\, d\boldsymbol{x} = \mathbf{0},$$

we have:

$$I_b = \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{\boldsymbol{x} - \phi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\Big\} \cdot \Big\{F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\Big\}\, d(\mathbb{P}_n - P_0)(\boldsymbol{x},t,\delta).$$

The negligibility of I_b now follows in the same way as (A.2.12), using the parametric nature of the function $\phi_{\boldsymbol{\alpha}}$ and the entropy properties of the class of functions

$$u \mapsto \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u) - F_{\hat{\boldsymbol{\alpha}}_n}(u).$$

The conclusion now follows from (A.2.13).

□

Remark A.2.1. *Note that the proof above yields the representation*

$$\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0 \sim n^{-1}\psi_{1\epsilon}'(\boldsymbol{\alpha}_0)^{-1}\sum_{i=1}^n(\boldsymbol{X}_i - \mathbb{E}(\boldsymbol{X}_i|T-\boldsymbol{\alpha}_0^T\boldsymbol{X}))\left\{\Delta_i - F_0(T_i - \boldsymbol{\alpha}_0^T\boldsymbol{X}_i\right\},$$

where $\psi_{1\epsilon}'(\boldsymbol{\alpha}_0)$ is given in Lemma 2.4.1.

A.3 Asymptotic behavior of the ESE

In this section we prove the asymptotic efficiency of the score estimator defined in Section 2.4.2. The proof of existence of the root and the consistency proof for the score estimator is similar to the proof of existence and consistency of the first score estimator defined in Section 2.4.1, thus omitted.

Proof of Theorem 2.4.2 (Asymptotic Normality). Since the proof is very similar to the proof of Theorem 2.4.1, we only give the main steps of the proof. As in the proof of Theorem 2.4.1, we can define $\psi_{2\epsilon,nh}$ at $\hat{\boldsymbol{\alpha}}_n$ by

$$\psi_{2\epsilon,nh}(\hat{\boldsymbol{\alpha}}_n) = \mathbf{0},$$

and $\psi_{2\epsilon,nh}(\hat{\boldsymbol{\alpha}}_n)$ is then a combination of one-sided limits at $\hat{\boldsymbol{\alpha}}_n$.
We prove that

$$\begin{aligned}&\psi_{2\epsilon,nh}(\hat{\boldsymbol{\alpha}}_n)\\ &= \int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \frac{\left\{\boldsymbol{x}f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - \varphi_{\boldsymbol{\alpha}_0}(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\right\}\left\{\delta - F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\right\}}{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\{1-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\}}\, d\mathbb{P}_n(t,x,\delta)\\ &\quad + \psi_{2\epsilon}'(\boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) + o_p\left(n^{-1/2} + (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)\right),\end{aligned} \tag{A.3.1}$$

where $\varphi_{\boldsymbol{\alpha}}$ is defined by

$$\varphi_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) = \mathbb{E}(\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}=t-\boldsymbol{\alpha}^T\boldsymbol{x})f_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}),$$

and $\psi_{2\epsilon}$ is defined by

$$\psi_{2\epsilon}(\boldsymbol{\alpha}) = \int_{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \frac{\Big\{\boldsymbol{x}f_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})-\varphi_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\Big\}\Big\{\delta-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\Big\}}{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\{1-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\}} \cdot dP_0(\boldsymbol{x},t,\delta).$$

Straightforward calculations show that

$$\begin{aligned}
\psi_{2\epsilon}'(\boldsymbol{\alpha}_0) &= \int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \frac{\big\{\boldsymbol{x}f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})-\varphi_{\boldsymbol{\alpha}_0}(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\big\}^2}{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\{1-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\}}\,dP_0(\boldsymbol{x},t,\delta)\\
&= \mathbb{E}_\epsilon\left\{\frac{f_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})^2\,\big\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\big\}\big\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\big\}'}{F_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\{1-F_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\}}\right\}\\
&= I_\epsilon(\boldsymbol{\alpha}_0).
\end{aligned}$$

(See also the derivation of the derivative ψ_ϵ' for the first score equation in the proof of Theorem 2.4.1, Part 1). Since

$$\sqrt{n}\int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \frac{\big\{\boldsymbol{x}f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})-\varphi_{\boldsymbol{\alpha}_0}(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\big\}\big\{\delta-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\big\}}{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\{1-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\}}\,d\mathbb{P}_n(\boldsymbol{x},t,\delta) \xrightarrow{d} N(\boldsymbol{0},I_\epsilon(\boldsymbol{\alpha}_0)),$$

(A.3.1) implies, using the nonsingularity of $\psi_{2\epsilon}'(\boldsymbol{\alpha}_0)$ and the consistency of $\hat{\boldsymbol{\alpha}}_n$,

$$\begin{aligned}
&\sqrt{n}(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)\\
&\quad= -\psi_{2\epsilon}'(\boldsymbol{\alpha}_0)^{-1}\Bigg\{\sqrt{n}\int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \frac{\boldsymbol{x}f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})-\varphi_{\boldsymbol{\alpha}_0}(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})}{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\{1-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\}}\\
&\qquad\qquad \cdot\big\{\delta-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\big\}\,d\mathbb{P}_n(\boldsymbol{x},t,\delta)\Bigg\}\\
&\qquad + o_p(1+\sqrt{n}(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0))\\
&\quad\xrightarrow{d} N\left(\boldsymbol{0},I_\epsilon(\boldsymbol{\alpha}_0)^{-1}\right).
\end{aligned}$$

Let, analogously to the start of the proof of Theorem 2.4.1, $\bar{\varphi}_{\hat{\boldsymbol{\alpha}}_n,\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}}$ be a (random) piecewise constant version of $\varphi_{\hat{\boldsymbol{\alpha}}_n}$, where, for a piecewise constant distribution function F with finitely many jumps at $\tau_1<\tau_2<\dots$, the function $\bar{\varphi}_{\boldsymbol{\alpha},F}$ is defined in the following way.

$$\bar{\varphi}_{\boldsymbol{\alpha},F}(u) = \begin{cases} \varphi_{\boldsymbol{\alpha}}(\tau_i), & \text{if } F_{\boldsymbol{\alpha}}(u) > F(\tau_i),\ u \in [\tau_i, \tau_{i+1}), \\ \varphi_{\boldsymbol{\alpha}}(s), & \text{if } F_{\boldsymbol{\alpha}}(u) = F(s),\ \text{for some } s \in [\tau_i, \tau_{i+1}), \\ \varphi_{\boldsymbol{\alpha}}(\tau_{i+1}), & \text{if } F_{\boldsymbol{\alpha}}(u) < F(\tau_i),\ u \in [\tau_i, \tau_{i+1}). \end{cases} \tag{A.3.2}$$

We now have:

$$\begin{aligned} \psi_{2\epsilon,nh}(\boldsymbol{\alpha}) &= \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \boldsymbol{x} f_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) \\ &\qquad \cdot \frac{\delta - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta) \\ &= \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\boldsymbol{x} f_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \varphi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} \\ &\qquad \cdot \frac{\delta - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta) \\ &\quad + \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\varphi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \bar{\varphi}_{n,\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} \\ &\qquad \cdot \frac{\delta - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta) \\ &= J + JJ. \end{aligned}$$

Let $\mathcal{F}$ be the set of piecewise constant distribution functions with finitely many jumps (like the MLE $\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}$), and let $\mathcal{K}_2$ be the set of functions

$$\begin{aligned} \mathcal{K}_2 = \Big\{ (\boldsymbol{x},t,\delta) \mapsto 1_{[\epsilon,1-\epsilon]}(F(t-\boldsymbol{\alpha}^T\boldsymbol{x}))\big\{\varphi_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - \bar{\varphi}_{\boldsymbol{\alpha},F}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big\} \\ \cdot \frac{\delta - F(t-\boldsymbol{\alpha}^T\boldsymbol{x}))}{F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\{1-F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\}} : F \in \mathcal{F},\ \boldsymbol{\alpha} \in \Theta \Bigg\}, \end{aligned}$$

where $\bar{\varphi}_{\boldsymbol{\alpha},F}$ is defined by (A.3.2). We add the function which is identically zero to $\mathcal{K}_2$. As in the proof of Theorem 2.4.1, the functions are uniformly bounded and also of uniformly bounded variation, using Assumption A4 and Assumption A5. For k_1 and k_2 in $\mathcal{K}_2$, we define:

$$d(k_1,k_2)^2 = \int \|k_1 - k_2\|^2 \, dP_0, \qquad k_1, k_2 \in \mathcal{K}_2.$$

For this distance, we therefore get, similarly as before, using Lemma A.1.1:

$$\sup_{\zeta>0} \zeta H_B\left(\zeta, \mathcal{K}_2, L_2(P_0)\right) = O(1),$$

which implies:

$$\int_0^\zeta H_B\left(u, \mathcal{K}_2, L_2(P_0)\right)^{1/2} du = O\left(\zeta^{1/2}\right), \qquad \zeta > 0.$$

Note that the indicator function keeps $F(t-\boldsymbol{\alpha}^T\boldsymbol{x})$ away from zero and one, which is essential for getting the bounded variation property.
Following the same steps as in the proof of Theorem 2.4.1, we get:

$$JJ = o_p\left(n^{-1/2} + \hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\right).$$

We now write

$$\begin{aligned}
J = &\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{\boldsymbol{x} f_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \varphi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\Big\}\\
&\qquad\cdot \frac{\delta - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta)\\
= &\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{\boldsymbol{x} f_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \varphi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\Big\}\\
&\qquad\cdot \frac{\delta - F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta)\\
&+\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{\boldsymbol{x} f_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \varphi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\Big\}\\
&\qquad\cdot \frac{F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{\hat{F}_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, dP_0(\boldsymbol{x},t,\delta)\\
&+\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{\boldsymbol{x} f_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \varphi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\Big\}\\
&\qquad\cdot \frac{F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{\hat{F}_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, d(\mathbb{P}_n - P_0)(\boldsymbol{x},t,\delta)\\
= &J_a + J_b + J_c.
\end{aligned}$$

For the term J_b we get:

$$\begin{aligned}
&\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{\boldsymbol{x} f_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \varphi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\Big\}\\
&\qquad\cdot \frac{F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{\hat{F}_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, dP_0(\boldsymbol{x},t,\delta)\\
&= \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X} = t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\right\} f_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\\
&\qquad\cdot \frac{F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{\hat{F}_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, dP_0(\boldsymbol{x},t,\delta)\\
&\quad+\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{f_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - f_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\Big\}\\
&\qquad\cdot \mathbb{E}(\boldsymbol{X}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X} = t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\frac{F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{\hat{F}_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, dP_0(\boldsymbol{x},t,\delta)
\end{aligned}$$

$$= \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{ f_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - f_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) \right\}$$

$$\cdot\, \mathbb{E}(\boldsymbol{X}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X} = t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) \frac{F_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{\hat{F}_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, dP_0(\boldsymbol{x},t,\delta)$$

$$= \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)\in[\epsilon,1-\epsilon]} \left\{ f_{nh,\hat{\boldsymbol{\alpha}}_n}(u) - f_{\hat{\boldsymbol{\alpha}}_n}(u) \right\} \mathbb{E}(\boldsymbol{X}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X} = u) \frac{F_{\hat{\boldsymbol{\alpha}}_n}(u) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)}{\hat{F}_{\hat{\boldsymbol{\alpha}}_n}(u)\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}u)\}}$$

$$\cdot\, f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(u)\, du.$$

Furthermore,

$$\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)\in[\epsilon,1-\epsilon]} \left\{ f_{nh,\hat{\boldsymbol{\alpha}}_n}(u) - f_{\hat{\boldsymbol{\alpha}}_n}(u) \right\} \mathbb{E}(\boldsymbol{X}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X} = u)$$

$$\cdot\, \frac{F_{\hat{\boldsymbol{\alpha}}_n}(u) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)}{\hat{F}_{\hat{\boldsymbol{\alpha}}_n}(u)\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}u)\}}\, f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(u)\, du$$

$$= h^{-2} \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)\in[\epsilon,1-\epsilon]} \left\{ \int K'((u-v)/h)\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(v)\, dv - f_{\hat{\boldsymbol{\alpha}}_n}(u) \right\}$$

$$\cdot\, \mathbb{E}(\boldsymbol{X}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X} = u) \frac{F_{\hat{\boldsymbol{\alpha}}_n}(u) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)}{\hat{F}_{\hat{\boldsymbol{\alpha}}_n}(u)\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)\}}\, f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(u)\, du$$

$$= h^{-2} \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)\in[\epsilon,1-\epsilon]} \left\{ \int K'((u-v)/h) \left\{ \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(v) - F_{\hat{\boldsymbol{\alpha}}_n}(v) \right\} dv \right\}$$

$$\cdot\, \mathbb{E}(\boldsymbol{X}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X} = u) \frac{F_{\hat{\boldsymbol{\alpha}}_n}(u) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)}{\hat{F}_{\hat{\boldsymbol{\alpha}}_n}(u)\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)\}}\, f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(u)\, du$$

$$+ \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)\in[\epsilon,1-\epsilon]} \left\{ \int K_h(u-v)\, dF_{\hat{\boldsymbol{\alpha}}_n}(v) - f_{\hat{\boldsymbol{\alpha}}_n}(u) \right\} \mathbb{E}(\boldsymbol{X}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X} = u)$$

$$\cdot\, \frac{F_{\hat{\boldsymbol{\alpha}}_n}(u) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)}{\hat{F}_{\hat{\boldsymbol{\alpha}}_n}(u)\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)\}}\, f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(u)\, du. \tag{A.3.3}$$

The last term on the right-hand side of the above equation has an upper bound of order $O_p(n^{-2/7-1/3}) = O_p(n^{-13/21}) = o_p(n^{-1/2})$, since

$$\left\{ \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)\in[\epsilon,1-\epsilon]} \left\{ \int K_h(u-v)\, dF_{\hat{\boldsymbol{\alpha}}_n}(v) - f_{\hat{\boldsymbol{\alpha}}_n}(u) \right\}^2 du \right\}^{1/2} = O_p\left(n^{-2/7}\right),$$

and

$$\left\{ \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u)\in[\epsilon,1-\epsilon]} \left\{ \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(u) - F_{\hat{\boldsymbol{\alpha}}_n}(u) \right\}^2 du \right\}^{1/2} = O_p\left(n^{-1/3}\right), \tag{A.3.4}$$

using Lemma 2.3.1 for the last relation. We also use the Cauchy-Schwarz inequality. The first term on the right of (A.3.3) is of order $O_p(n^{1/7-2/3}) = O_p(n^{-11/21}) =$

$o_p(n^{-1/2})$ by (A.3.4) and using

$$\left|h^{-2}\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n}(u)\in[\epsilon,1-\epsilon]}\left[\int K'((u-v)/h)\left\{\hat F_{n,\hat{\boldsymbol\alpha}_n}(v)-F_{\hat{\boldsymbol\alpha}_n}(v)\right\}dv\right]\right.$$
$$\left.\cdot\mathbb{E}(\boldsymbol X|T-\hat{\boldsymbol\alpha}_n^T\boldsymbol X=u)\frac{F_{\hat{\boldsymbol\alpha}_n}(u)-\hat F_{n,\hat{\boldsymbol\alpha}_n}(u)}{\hat F_{\hat{\boldsymbol\alpha}_n}(u)\{1-\hat F_{n,\hat{\boldsymbol\alpha}_n}(u)\}}f_{T-\hat{\boldsymbol\alpha}_n^T\boldsymbol X}(u)\,du\right|$$
$$=h^{-1}\left|\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n}(u)\in[\epsilon,1-\epsilon]}\left[\int K'(w)\left\{\hat F_{n,\hat{\boldsymbol\alpha}_n}(u-hw)-F_{\hat{\boldsymbol\alpha}_n}(u-hw)\right\}dw\right]\right.$$
$$\left.\cdot\mathbb{E}(\boldsymbol X|T-\hat{\boldsymbol\alpha}_n^T\boldsymbol X=u)\frac{F_{\hat{\boldsymbol\alpha}_n}(u)-\hat F_{n,\hat{\boldsymbol\alpha}_n}(u)}{\hat F_{\hat{\boldsymbol\alpha}_n}(u)\{1-\hat F_{n,\hat{\boldsymbol\alpha}_n}(u)\}}f_{T-\hat{\boldsymbol\alpha}_n^T\boldsymbol X}(u)\,du\right|$$
$$\le ch^{-1}\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n}(u)\in[\epsilon/2,1-\epsilon/2]}\left\{\hat F_{n,\hat{\boldsymbol\alpha}_n}(u)-F_{\hat{\boldsymbol\alpha}_n}(u)\right\}^2du,$$

for small h and a constant $c>0$, where we first use Fubini's theorem and next the Cauchy-Schwarz inequality in the last inequality, together with

$$\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n}(u)\in[\epsilon,1-\epsilon]}\left\{\hat F_{n,\hat{\boldsymbol\alpha}_n}(u-hw)-F_{\hat{\boldsymbol\alpha}_n}(u-hw)\right\}^2du$$
$$\le\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n}(u)\in[\epsilon/2,1-\epsilon/2]}\left\{\hat F_{n,\hat{\boldsymbol\alpha}_n}(u)-F_{\hat{\boldsymbol\alpha}_n}(u)\right\}^2du,$$

for small $h>0$, together with $w\in[-1,1]$. Finally we use Lemma 2.3.1 again. Note that a bandwidth choice $h\asymp n^{-1/5}$ corresponds to the order $O_p(n^{1/5-2/3})=O_p(n^{-7/15})$ for the first term on the right hand side of (A.3.3). Consequently, when the bandwidth is of order $n^{-1/5}$, this term is no longer negligible in the proofs and therefore we use a bandwidth $h\asymp n^{-1/7}$ in Theorem 2.4.2. The latter bandwidth choice moreover corresponds to the classical choice for estimating the density in the current status model (see e.g. Groeneboom et al., 2010).

For the term J_c we argue similarly as before using Lemma A.0.1 that

$$J_c=o_p\left(n^{-1/2}\right).$$

Finally,

$$J_a=\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\in[\epsilon,1-\epsilon]}\left\{\boldsymbol x f_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)-\varphi_{\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\right\}$$
$$\cdot\frac{\delta-\hat F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)}{\hat F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\{1-\hat F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\}}\,d\big(\mathbb{P}_n-P_0\big)(\boldsymbol x,t,\delta)$$
$$+\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\in[\epsilon,1-\epsilon]}\left\{\boldsymbol x f_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)-\varphi_{\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\right\}$$
$$\cdot\frac{F_0(t-\boldsymbol\alpha_0^T\boldsymbol x)-\hat F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)}{\hat F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\{1-\hat F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol x)\}}\,dP_0(\boldsymbol x,t,\delta).$$

This time we consider the class of functions

$$\mathcal{K}_2' = \Big\{(\boldsymbol{x},t,\delta) \mapsto 1_{[\epsilon,1-\epsilon]}(F(t-\boldsymbol{\alpha}^T\boldsymbol{x}))\big(\boldsymbol{x}f(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - \varphi_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big) \cdot \frac{\delta - F(t-\boldsymbol{\alpha}^T\boldsymbol{x})}{F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\{1-F(t-\boldsymbol{\alpha}^T\boldsymbol{x})\}} : F \in \mathcal{F},\ f \in \mathcal{F}',\ \boldsymbol{\alpha} \in \Theta\Big\},$$

where $\mathcal{F}'$ is a class of uniformly bounded functions of uniformly bounded variation (which have the interpretation of estimates of $F_{\boldsymbol{\alpha}}'$), to which we add the function

$$(\boldsymbol{x},t,\delta) \mapsto 1_{[\epsilon,1-\epsilon]}(F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}))\big(\boldsymbol{x}f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - \varphi_{\boldsymbol{\alpha}_0}(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\big) \cdot \frac{\delta - F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})}{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\{1-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\}}.$$

So we get, using Lemma A.1.1,

$$\sup_{\zeta>0} \zeta H_B\left(\zeta, \mathcal{K}_2', L_2(P_0)\right) = O(1),$$

which implies:

$$\int_0^\zeta H_B\left(u, \mathcal{K}_2', L_2(P_0)\right)^{1/2}\, du = O\left(\zeta^{1/2}\right), \qquad \zeta > 0.$$

As before, we now get:

$$\begin{aligned}
&\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{\boldsymbol{x}f_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \varphi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\Big\} \\
&\qquad \cdot \frac{\delta - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x},t,\delta) \\
&= \int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{\boldsymbol{x}f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - \varphi_{\boldsymbol{\alpha}_0}(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\Big\} \\
&\qquad \cdot \frac{\delta - F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})}{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\{1-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\}}\, d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x},t,\delta) \\
&\quad + o_p\left(n^{-1/2} + \hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\right)
\end{aligned}$$

and

$$\begin{aligned}
&\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{\boldsymbol{x}f_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - \varphi_{\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\Big\} \\
&\qquad \cdot \frac{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1-\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, dP_0(\boldsymbol{x},t,\delta) \\
&= \Bigg\{\int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \Big\{\boldsymbol{x}f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - \varphi_{\boldsymbol{\alpha}_0}(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\Big\} \\
&\qquad \cdot \frac{f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})x'}{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\{1-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\}}\, dP_0(\boldsymbol{x},t,\delta)\Bigg\}\left(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\right) \\
&\quad + o_p\left(n^{-1/2} + \hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\right).
\end{aligned}$$

The result now follows. □

A.4 Asymptotic behavior of the plug-in estimator

In this section we first sketch in Section A.4.1 the proof of consistency of the plug-in estimator, denoted by $\hat{\boldsymbol{\alpha}}_n$. This is the second result stated in Theorem 2.4.3. The proof of existence of a root is similar to the proof of existence of a root of the SSE and omitted. We next prove the asymptotic normality result of the plug-in estimator, which is the third result given in Theorem 2.4.3. The proof of Theorem 2.4.5 on the asymptotic representation of the plug-in estimator as a sum of i.i.d. random variables follows from the proof of 2.4.3. The asymptotic distribution of the estimator of the intercept, given in Theorem 2.5.1, is proved in Section A.4.2.

Before we start the proofs, we give, in Lemma A.4.1, some auxiliary results on the L_2-distance between the plug-in estimate $F_{nh,\boldsymbol{\alpha}}$ and $F_{\boldsymbol{\alpha}}$ and between the partial derivative of the plug-in estimate $\partial_{\boldsymbol{\alpha}} F_{nh,\boldsymbol{\alpha}}$ and $\partial_{\boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}$. Next, we follow the arguments used to prove the asymptotic normality of the SSE and ESE and give a similar proof for the limiting distribution of the plug-in estimator.

Lemma A.4.1. *Let the conditions of Theorem 2.4.3 be satisfied. Then we have, for the estimate $F_{nh,\boldsymbol{\alpha}}$, defined by (2.4.6) that*

$$\int_{F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\}^2 \, dG(\boldsymbol{x},t) \tag{A.4.1}$$
$$= O_p\left(\frac{1}{nh}\right) + O_p\left(h^4\right),$$

$$\int_{F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\partial_{\boldsymbol{\alpha}} F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - \partial_{\boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\}^2 \, dG(\boldsymbol{x},t) \tag{A.4.2}$$
$$= O_p\left(\frac{1}{nh^3}\right) + O_p\left(h^2\right)$$

uniformly in $\boldsymbol{\alpha} \in \Theta$.

proof of Lemma A.4.1. We first prove the first part and show that (A.4.1) holds. Recall that

$$F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) = \frac{g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}{g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})},$$

where

$$g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) = \int \delta K_h(t-\boldsymbol{\alpha}^T\boldsymbol{x} - u + \boldsymbol{\alpha}^T\boldsymbol{y}) \, d\mathbb{P}_n(\boldsymbol{y},u,\delta),$$

and

$$g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) = \int K_h(t-\boldsymbol{\alpha}^T\boldsymbol{x}-u+\boldsymbol{\alpha}^T\boldsymbol{y})\,d\mathbb{P}_n(\boldsymbol{y},u,\delta).$$

Moreover,

$$F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) = \int F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T(\boldsymbol{y}-\boldsymbol{x}))f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{y}|t-\boldsymbol{\alpha}^T\boldsymbol{x})\,d\boldsymbol{y}.$$

We first investigate the bias part:

$$\begin{aligned}
&\mathbb{E}g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) = \int F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})K_h(t-\boldsymbol{\alpha}^T\boldsymbol{x}-u+\boldsymbol{\alpha}^T\boldsymbol{y})\,dG(\boldsymbol{y},u)\\
&= \int F_0(v+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{y})K_h(t-\boldsymbol{\alpha}^T\boldsymbol{x}-v)\,f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(v)\,f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{y}|v)\,d\boldsymbol{y}\,dv\\
&= \int F_0(t-\boldsymbol{\alpha}^T\boldsymbol{x}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T\boldsymbol{y}-hw)K(w)\\
&\qquad\qquad \cdot f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}-hw)\,f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{y}|t-\boldsymbol{\alpha}^T\boldsymbol{x}-hw)\,d\boldsymbol{y}\,dw\\
&= f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\int F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T(\boldsymbol{y}-\boldsymbol{x}))f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{y}|t-\boldsymbol{\alpha}^T\boldsymbol{x})\,d\boldsymbol{y}\\
&\qquad + O\left(h^2\right),
\end{aligned}$$

uniformly in $\boldsymbol{\alpha}\in\Theta$ and $\boldsymbol{x},t$ varying over a bounded set, due to the assumptions of Theorem 2.4.3. In a similar way, we get

$$\mathbb{E}g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) = f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})+O\left(h^2\right),$$

uniformly in $\boldsymbol{\alpha}\in\Theta$ and $\boldsymbol{x},t$ varying over a bounded set. So we find:

$$\frac{\mathbb{E}g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}{\mathbb{E}g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})} = F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})+O\left(h^2\right),$$

uniformly in $\boldsymbol{\alpha}\in\Theta$ and $\boldsymbol{x},t$ varying over a bounded set, such that $\mathbb{E}g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})$ stays away from zero. We now get:

$$\begin{aligned}
F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) &= \frac{g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})-\mathbb{E}g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}{g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}\\
&+\mathbb{E}g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\frac{\mathbb{E}g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})-g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}{g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\mathbb{E}g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}+O\left(h^2\right),
\end{aligned}$$

and

$$\begin{aligned}
\left\{F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\}^2 &\le 3\left\{\frac{g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})-\mathbb{E}g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}{g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}\right\}^2\\
+3\left\{\mathbb{E}g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\frac{\mathbb{E}g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})-g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}{g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\mathbb{E}g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}\right\}^2 &+O\left(h^4\right),
\end{aligned} \tag{A.4.3}$$

uniformly in $\boldsymbol{\alpha} \in \Theta$ and $\boldsymbol{x}, t$ varying over a bounded set, such that $\mathbb{E}g_{nh,1,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})$ stays away from zero.

Since $\eta > 0$ is chosen in such a way that $a_1(\beta) = F_{\boldsymbol{\alpha}}^{-1}(\epsilon) > a$, $b_1(\beta) = F_{\boldsymbol{\alpha}}^{-1}(1-\epsilon) < b$, for each $\boldsymbol{\alpha} \in \Theta$ and since $g_{nh,\boldsymbol{\alpha}}$ stays away from zero with probability tending to one if $\epsilon < F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) < 1 - \epsilon$ we get:

$$\int_{F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\frac{g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - \mathbb{E}g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}{g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}\right\}^2 dG(\boldsymbol{x},t)$$
$$\lesssim \int_{F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - \mathbb{E}g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\}^2 dG(\boldsymbol{x},t).$$

Furthermore

$$\mathbb{E}\left\{g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - \mathbb{E}g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\}^2$$
$$= \mathbb{E}\left\{\int \delta K_h(t-\boldsymbol{\alpha}^T\boldsymbol{x} - u + \boldsymbol{\alpha}^T\boldsymbol{y})\, d(\mathbb{P}_n - P_0)(\boldsymbol{y},u,\delta)\right\}^2 = O\left(\frac{1}{nh}\right),$$

uniformly for $(\boldsymbol{x}, t)$ in a bounded region, so we get:

$$\mathbb{E}\int_{F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - \mathbb{E}g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\}^2 dG(\boldsymbol{x},t) = O\left(\frac{1}{nh}\right).$$

Hence

$$\int_{F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\frac{g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - \mathbb{E}g_{nh,1,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}{g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}\right\}^2 dG(\boldsymbol{x},t) = O_p\left(\frac{1}{nh}\right).$$

The second term on the right-hand side of (A.4.3) can be treated in a similar way. This proves (A.4.1).

We next continue with the proof of (A.4.2). We have:

$$\partial_{\boldsymbol{\alpha}} F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) \tag{A.4.4}$$
$$= \frac{\int(\boldsymbol{y}-\boldsymbol{x})\{\delta - F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\}K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x} - u + \boldsymbol{\alpha}^T\boldsymbol{y})\, d\mathbb{P}_n(\boldsymbol{y},u,\delta)}{g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}.$$

We consider the numerator of (A.4.4). It can be rewritten as

$$\int(\boldsymbol{y}-\boldsymbol{x})\{\delta - F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})\}K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x} - u + \boldsymbol{\alpha}^T\boldsymbol{y})\, d\mathbb{P}_n(\boldsymbol{y},u,\delta)$$
$$+ \int(\boldsymbol{y}-\boldsymbol{x})\{F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y}) - F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\}K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x} - u + \boldsymbol{\alpha}^T\boldsymbol{y})\, d\mathbb{G}_n(\boldsymbol{y},u)$$
$$+ \{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\}\int(\boldsymbol{y}-\boldsymbol{x})K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x} - u + \boldsymbol{\alpha}^T\boldsymbol{y})\, d\mathbb{G}_n(\boldsymbol{y},u).$$

The first term can be written as

$$A_n(\boldsymbol{x},t,\beta) \stackrel{\text{def}}{=} \int(\boldsymbol{y}-\boldsymbol{x})\{\delta - F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})\}K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x} - u + \boldsymbol{\alpha}^T\boldsymbol{y})\, d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{y},u,\delta),$$

and we have for each component $A_{n,j}$ of $A_n, 1 \le j \le d$:

$$\mathbb{E}\int_{F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} A_{n,j}(\boldsymbol{x},t,\boldsymbol{\alpha})^2\,dG(\boldsymbol{x},t) \le \mathbb{E}\int A_{n,j}(\boldsymbol{x},t,\boldsymbol{\alpha})^2\,dG(\boldsymbol{x},t)$$
$$\sim \frac{1}{nh^3}\int \mathsf{var}(X_j|v)F_0(v)\{1-F_0(v)\}f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(v)\,dv\int K'(u)^2\,du,\ n\to\infty.$$

In the second term we must compare $F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})$ with

$$F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) = \int F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T(\boldsymbol{z}-\boldsymbol{x}))f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{z}|t-\boldsymbol{\alpha}^T\boldsymbol{x})\,d\boldsymbol{z}.$$

We can write:

$$F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})$$
$$=\int\left\{F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T(\boldsymbol{z}-\boldsymbol{x}))\right\}f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{z}|t-\boldsymbol{\alpha}^T\boldsymbol{x})\,d\boldsymbol{z}.$$

So we find for the second term:

$$\begin{aligned}
&B_n(\boldsymbol{x},t,\boldsymbol{\alpha})\\
&\stackrel{\mathsf{def}}{=}\int(\boldsymbol{y}-\boldsymbol{x})\left\{F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\}K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x}-u+\boldsymbol{\alpha}^T\boldsymbol{y})\,d\mathbb{G}_n(\boldsymbol{y},u)\\
&=\int\int(\boldsymbol{y}-\boldsymbol{x})\left\{F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T(\boldsymbol{z}-\boldsymbol{x}))\right\}\\
&\qquad\cdot f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{z}|t-\boldsymbol{\alpha}^T\boldsymbol{x})\,d\boldsymbol{z}K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x}-u+\boldsymbol{\alpha}^T\boldsymbol{y})\,d\mathbb{G}_n(\boldsymbol{y},u)\\
&=\int(\boldsymbol{y}-\boldsymbol{x})\int\left\{F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T(\boldsymbol{z}-\boldsymbol{x}))\right\}\\
&\qquad\cdot f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{z}|t-\boldsymbol{\alpha}^T\boldsymbol{x})\,d\boldsymbol{z}K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x}-u+\boldsymbol{\alpha}^T\boldsymbol{y})\,dG(\boldsymbol{y},u)\\
&\quad+\int(\boldsymbol{y}-\boldsymbol{x})\int\left\{F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T(\boldsymbol{z}-\boldsymbol{x}))\right\}\\
&\qquad\cdot f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{z}|t-\boldsymbol{\alpha}^T\boldsymbol{x})\,d\boldsymbol{z}K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x}-u+\boldsymbol{\alpha}^T\boldsymbol{y})\,d\big(\mathbb{G}_n-G\big)(\boldsymbol{y},u)\\
&=f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\partial_{\boldsymbol{\alpha}}F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})+O(h)+O_p\left(\frac{1}{nh^3}\right).
\end{aligned}$$

where, using integration by parts, the last line follows by straightforward calculation. Since

$$g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) = f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})+O_p(h^2),$$

we get:

$$\int_{F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]}\left\|\frac{B_n(\boldsymbol{x},t,\boldsymbol{\alpha})}{g_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})}-\partial_{\boldsymbol{\alpha}}F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\|^2 dG(\boldsymbol{x},t)$$
$$=O_p\left(\frac{1}{nh^3}\right)+O_p\left(h^2\right).$$

Finally, defining

$$C_n(\boldsymbol{x},t,\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\} \int (\boldsymbol{y}-\boldsymbol{x})K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x}-u+\boldsymbol{\alpha}^T\boldsymbol{y})\,d\mathbb{G}_n(\boldsymbol{y},u),$$

we get, using,

$$\begin{aligned}
&\int (\boldsymbol{y}-\boldsymbol{x})K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x}-u+\boldsymbol{\alpha}^T\boldsymbol{y})\,d\mathbb{G}_n(\boldsymbol{y},u)\\
&=\int (\boldsymbol{y}-\boldsymbol{x})K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x}-u+\boldsymbol{\alpha}^T\boldsymbol{y})\,dG(\boldsymbol{y},u)\\
&\quad+\int (\boldsymbol{y}-\boldsymbol{x})K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x}-u+\boldsymbol{\alpha}^T\boldsymbol{y})\,d(\mathbb{G}_n-G)(\boldsymbol{y},u)\\
&=\int (\boldsymbol{y}-\boldsymbol{x})K_h'(t-\boldsymbol{\alpha}^T\boldsymbol{x}-v)f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(v)f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{y}|v)\,dv\,d\boldsymbol{y}+O_p\left(\frac{1}{nh^3}\right)\\
&=\int (\boldsymbol{y}-\boldsymbol{x})K_h(t-\boldsymbol{\alpha}^T\boldsymbol{x}-v)\frac{d}{dv}\left\{f_{T-\boldsymbol{\alpha}^T\boldsymbol{X}}(v)f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{y}|v)\right\}\,dv\,d\boldsymbol{y}+O_p\left(\frac{1}{nh^3}\right)\\
&=O_p(1),
\end{aligned}$$

and using the first part of Lemma A.4.1 for the factor $F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})$ that

$$\int_{F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \|C_n(\boldsymbol{x},t,\boldsymbol{\alpha})\|^2\,dG(\boldsymbol{x},t) = O_p\left(\frac{1}{nh}\right)+O_p\left(h^4\right).$$

This proves (A.4.2). □

A.4.1 Consistency and asymptotic normality of the plug-in estimator

We first prove that $\hat{\boldsymbol{\alpha}}_n$ is a consistent estimate of $\boldsymbol{\alpha}_0$.

Proof of Theorem 2.4.3, Part 1 (Consistency). We assume that $\hat{\boldsymbol{\alpha}}_n$ is contained in the compact set Θ, and hence the sequence $(\hat{\boldsymbol{\alpha}}_n)$ has a subsequence $(\hat{\boldsymbol{\alpha}}_{n_k} = \hat{\boldsymbol{\alpha}}_{n_k}(\omega))$, converging to an element $\boldsymbol{\alpha}_*$. It is easily seen that, if $\hat{\boldsymbol{\alpha}}_{n_k} = \hat{\boldsymbol{\alpha}}_{n_k}(\omega) \longrightarrow \boldsymbol{\alpha}_*$, we get:

$$\begin{aligned}
F_{n_kh,\hat{\boldsymbol{\alpha}}_{n_k}}(t-\hat{\boldsymbol{\alpha}}_{n_k}^T\boldsymbol{x}) \longrightarrow & F_{\boldsymbol{\alpha}_*}(t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\\
&\stackrel{\text{def}}{=} \int F_0(t-\boldsymbol{\alpha}_*^T\boldsymbol{x}+(\boldsymbol{\alpha}_*-\boldsymbol{\alpha}_0)^T\boldsymbol{y})f_{\boldsymbol{X}|T-\boldsymbol{\alpha}_*^T\boldsymbol{X}}(\boldsymbol{y}|t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\,d\boldsymbol{y}.
\end{aligned}$$

In the limit we get the relation:

$$\lim_{k\to\infty} -(\hat{\boldsymbol{\alpha}}_{n_k}-\boldsymbol{\alpha}_0)^T \int_{F_{n_k h,\hat{\boldsymbol{\alpha}}_{n_k}}(t-\hat{\boldsymbol{\alpha}}_{n_k}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \frac{\{\delta - F_{n_k h,\hat{\boldsymbol{\alpha}}_{n_k}}(t-\hat{\boldsymbol{\alpha}}_{n_k}^T\boldsymbol{x})\}\partial_{\boldsymbol{\alpha}} F_{n_k h,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\big|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_{n_k}}}{F_{n_k,\hat{\boldsymbol{\alpha}}_{n_k}}(t-\hat{\boldsymbol{\alpha}}_{n_k}^T\boldsymbol{x})\{1-F_{n_k,\hat{\boldsymbol{\alpha}}_{n_k}}(t-\hat{\boldsymbol{\alpha}}_{n_k}^T\boldsymbol{x})\}},\, d\mathbb{P}_{n_k}(\boldsymbol{x},t,\delta)$$
$$= -(\boldsymbol{\alpha}_*-\boldsymbol{\alpha}_0)^T \int_{F_{\boldsymbol{\alpha}_*}(t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \frac{\{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) - F_{\boldsymbol{\alpha}_*}(t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\}\partial_{\boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_*}}{F_{\boldsymbol{\alpha}_*}(t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\{1-F_{\boldsymbol{\alpha}_*}(t-\boldsymbol{\alpha}_*^T\boldsymbol{x})\}}\, dG(\boldsymbol{x},t)=0,$$

which can only mean $\boldsymbol{\alpha}_*=\boldsymbol{\alpha}_0$ by condition (2.4.9).

□

We next continue with the proof of the asymptotic normality of the plug-in estimator.

Proof of Theorem 2.4.3, Part 2 (Asymptotic Normality). To prove the asymptotic normality of the plug-in estimator, we follow the reasoning of the corresponding proofs for the SSE and ESE. We prove that

$$\psi_{3\epsilon,nh}(\hat{\boldsymbol{\alpha}}_n) = \int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{E(\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X}=t-\boldsymbol{\alpha}_0^T\boldsymbol{x})-\boldsymbol{x}\right\} f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}) \cdot \frac{\{\delta-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\}}{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\{1-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\}}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta)$$
$$+\psi_{3\epsilon}'(\boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)+o_p\left(n^{-1/2}+(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)\right),$$

where $\psi_{3\epsilon}$ is defined by

$$\psi_{3\epsilon}(\boldsymbol{\alpha}) = \int_{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \frac{\partial_{\boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\Big\{\delta - F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\Big\}}{F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\{1-F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\}}\, dP_0(\boldsymbol{x},t,\delta),$$

and

$$\psi_{3\epsilon}'(\boldsymbol{\alpha}_0) = -\mathbb{E}_\epsilon\left\{\frac{f_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})^2\,\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T}{F_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\{1-F_0(T-\boldsymbol{\alpha}_0^T\boldsymbol{X})\}}\right\}$$
$$= -I_\epsilon(\boldsymbol{\alpha}_0),$$

which follows by straightforward calculations after noting that

$$\partial_{\boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})$$
$$=\int(\boldsymbol{y}-\boldsymbol{x})f_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T(\boldsymbol{y}-\boldsymbol{x}))f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{y}|T-\boldsymbol{\alpha}^T\boldsymbol{X}=t-\boldsymbol{\alpha}^T\boldsymbol{x})d\boldsymbol{y}$$
$$+\int F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x}+(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)^T(\boldsymbol{y}-\boldsymbol{x}))\partial_{\boldsymbol{\alpha}} f_{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X}}(\boldsymbol{y}|T-\boldsymbol{\alpha}^T\boldsymbol{X}=t-\boldsymbol{\alpha}^T\boldsymbol{x})\, dG(\boldsymbol{x},t),$$

which is, at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$, equal to

$$f_0(t - \boldsymbol{\alpha}_0^T \boldsymbol{x}) \mathbb{E}\left\{\boldsymbol{X} - \boldsymbol{x} | T - \boldsymbol{\alpha}_0^T \boldsymbol{X} = t - \boldsymbol{\alpha}_0^T \boldsymbol{x}\right\}.$$

We have:

$$\begin{aligned}
\psi_{3\epsilon,nh}(\hat{\boldsymbol{\alpha}}_n) = \int_{F_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} & \partial_{\boldsymbol{\alpha}} F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \mid_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_n} \\
& \cdot \frac{\delta - F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1 - F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, d\mathbb{P}_n(\boldsymbol{x}, t, \delta) \\
= \int_{F_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} & \partial_{\boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \mid_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_n} \\
& \cdot \frac{\delta - F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1 - F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, d\mathbb{P}_n(\boldsymbol{x}, t, \delta) \\
+ \int_{F_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} & \left\{\partial_{\boldsymbol{\alpha}} F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \mid_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_n} - \partial_{\boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \mid_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_n}\right\} \\
& \cdot \frac{\delta - F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1 - F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\, d\mathbb{P}_n(\boldsymbol{x}, t, \delta) \\
= L + LL. &
\end{aligned}$$

Let $\mathcal{F}$ be a class of functions with the property that

$$\int_{\epsilon/2 < F_{\boldsymbol{\alpha}}(u) < 1-\epsilon/2} f'(u)^2\, du \le M.$$

if $f \in \mathcal{F}$, for a fixed $M > 0$. Using Proposition 5.1.9, p. 393 in Giné and Nickl (2015), with $m = 1$, $p = 2$ and $h \asymp n^{-1/5}$, we may assume that the functions $u \to F_{nh,\boldsymbol{\alpha}}(u)$ and $u \to \partial_{\boldsymbol{\alpha}} F_{nh,\boldsymbol{\alpha}}(u)$ belong to $\mathcal{F}$. Since the plug-in estimates are monotonically increasing with probability tending to one we get that the function

$$(\boldsymbol{x}, t) \mapsto 1_{[\epsilon,1-\epsilon]}\left(F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\right),$$

can be written in the form

$$\begin{aligned}
(\boldsymbol{x}, t) \mapsto & 1_{\left[a_{\epsilon,F_{nh,\boldsymbol{\alpha}}}, b_{\epsilon,F_{nh,\boldsymbol{\alpha}}}\right]}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \\
& = 1_{\left[a_{\epsilon,F_{nh,\boldsymbol{\alpha}}}, \infty\right)}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) - 1_{\left(b_{\epsilon,F_{nh,\boldsymbol{\alpha}}}, \infty\right)}(t - \boldsymbol{\alpha}^T\boldsymbol{x}),
\end{aligned}$$

for $a_{\epsilon,F_{nh,\boldsymbol{\alpha}}} \le b_{\epsilon,F_{nh,\boldsymbol{\alpha}}}$ for large n, with probability tending to one. The function is therefore of uniformly bounded variation for n sufficiently large (see also the proofs of Theorems 2.4.1 and 2.4.2). It now follows that the bracketing $\zeta-$entropy $H_B(\zeta, \mathcal{K}_3, L_2(P_0))$ for the class $\mathcal{K}_3$ of functions consisting of the function which is identically zero and the functions

$$\mathcal{K}_3 = \Bigg\{(\boldsymbol{x}, t, \delta) \mapsto \left\{\partial_{\boldsymbol{\alpha}} F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) - \partial_{\boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\right\} \cdot \frac{\delta - F(t - \boldsymbol{\alpha}^T\boldsymbol{x})}{F(t - \boldsymbol{\alpha}^T\boldsymbol{x})\{1 - F(t - \boldsymbol{\alpha}^T\boldsymbol{x})\}} 1_{[\epsilon,1-\epsilon]}(F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})) : F \in \mathcal{F},\, \boldsymbol{\alpha} \in \Theta\Bigg\},$$

w.r.t. the L_2-distance satisfies:

$$\sup_{\zeta > 0} \zeta H_B\left(\zeta, \mathcal{K}_3, L_2(P_0)\right) = O(1),$$

which implies:

$$\int_0^{\zeta} H_B\left(u, \mathcal{K}_3, L_2(P_0)\right)^{1/2} du = O\left(\zeta^{1/2}\right), \qquad \zeta > 0.$$

Moreover, by Lemma A.4.1 we also have:

$$\int_{F_{nh,\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x}) \in [\epsilon,1-\epsilon]} \Bigg\{\left\{\partial_{\boldsymbol{\alpha}} F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) - \partial_{\boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\right\} \cdot \frac{\delta - F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})}{F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\{1 - F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x})\}}\Bigg\}^2 dP_0(\boldsymbol{x}, t, \delta) \xrightarrow{p} 0.$$

This implies, by an application of Lemma A.0.1, that

$$\int_{F_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) \in [\epsilon,1-\epsilon]} \Bigg\{\left\{\partial_{\boldsymbol{\alpha}} F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \mid_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_n} - \partial_{\boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \mid_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_n}\right\} \cdot \frac{\delta - F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{F_{nh,\hat{\boldsymbol{\alpha}}_n b}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1 - F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\Bigg\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, t, \delta)$$
$$= o_p\left(n^{-1/2}\right).$$

Furthermore, an application of the Cauchy-Schwarz inequality and Lemma A.4.1 yield that

$$\sqrt{n} \int_{F_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\beta x) \in [\epsilon,1-\epsilon]} \left\{\partial_{\boldsymbol{\alpha}} F_{\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \mid_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_n} - \partial_{\boldsymbol{\alpha}} F_{nh,\boldsymbol{\alpha}}(t - \boldsymbol{\alpha}^T\boldsymbol{x}) \mid_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}_n}\right\} \cdot \left\{\frac{F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x}) - F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\{1 - F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\}}\right\} dP_0(\boldsymbol{x}, t, \delta)$$
$$= O_p\left(n^{-1/10}\right) + o_p\left(\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)\right).$$

We conclude that

$$LL = o_p\left(n^{-1/2} + (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)\right).$$

We now write:

$$
\begin{aligned}
L &= \int_{F_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \partial_{\boldsymbol\alpha} F_{\boldsymbol\alpha}(t-\boldsymbol\alpha^T\boldsymbol{x})\mid_{\boldsymbol\alpha=\hat{\boldsymbol\alpha}_n}\\
&\qquad\qquad \cdot \frac{\delta - F_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})}{F_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\{1-F_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\}}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta)\\
&= \int_{F_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \partial_{\boldsymbol\alpha} F_{\boldsymbol\alpha}(t-\boldsymbol\alpha^T\boldsymbol{x})\mid_{\boldsymbol\alpha=\hat{\boldsymbol\alpha}_n}\\
&\qquad\qquad \cdot \frac{\delta - F_{\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})}{F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\{1-F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\}}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta)\\
&\quad + \int_{F_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \partial_{\boldsymbol\alpha} F_{\boldsymbol\alpha}(t-\boldsymbol\alpha^T\boldsymbol{x})\mid_{\boldsymbol\alpha=\hat{\boldsymbol\alpha}_n}\\
&\qquad\qquad\qquad \cdot \frac{F_{\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x}) - F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})}{\hat{F}_{\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\{1-F_{n,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\}}\, d\mathbb{P}_n(\boldsymbol{x},t,\delta)\\
&= L_a + L_b.
\end{aligned}
$$

We now get, using Lemma A.4.1 and

$$
\partial_{\boldsymbol\alpha} F_{\boldsymbol\alpha}(t-\boldsymbol\alpha^T\boldsymbol{x})\mid_{\boldsymbol\alpha=\hat{\boldsymbol\alpha}_n} = \mathbb{E}(\boldsymbol{X}-\boldsymbol{x}|T-\hat{\boldsymbol\alpha}_n^T\boldsymbol{X} = t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x}) f_0(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x}) + O_p\left(\hat{\boldsymbol\alpha}_n-\boldsymbol\alpha_0\right),
$$

that

$$
L_b = o_p\left(n^{-1/2}+\hat{\boldsymbol\alpha}_n-\boldsymbol\alpha_0\right).
$$

The result of Theorem 2.4.3 now follows by showing that

$$
\begin{aligned}
&\int_{F_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \partial_{\boldsymbol\alpha} F_{\boldsymbol\alpha}(t-\boldsymbol\alpha^T\boldsymbol{x})\mid_{\boldsymbol\alpha=\hat{\boldsymbol\alpha}_n}\\
&\qquad\qquad \cdot \frac{\delta - F_{\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})}{F_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\{1-F_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\}}\, d(\mathbb{P}_n-P_0)(\boldsymbol{x},t,\delta)\\
&= \int_{F_0(t-\boldsymbol\alpha_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \partial_{\boldsymbol\alpha} F_{\boldsymbol\alpha}(t-\boldsymbol\alpha^T\boldsymbol{x})\mid_{\boldsymbol\alpha=\boldsymbol\alpha_0}\\
&\qquad\qquad \cdot \frac{\delta - F_0(t-\boldsymbol\alpha_0^T\boldsymbol{x})}{F_0(t-\boldsymbol\alpha_0^T\boldsymbol{x})\{1-F_0(t-\boldsymbol\alpha_0^T\boldsymbol{x})\}}\, d(\mathbb{P}_n-P_0)(\boldsymbol{x},t,\delta)\\
&\quad + o_p\left(n^{-1/2}+\hat{\boldsymbol\alpha}_n-\boldsymbol\alpha_0\right), \qquad \text{(A.4.5)}
\end{aligned}
$$

and,

$$
\begin{aligned}
&\int_{F_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \partial_{\boldsymbol\alpha} F_{\boldsymbol\alpha}(t-\boldsymbol\alpha^T\boldsymbol{x})\mid_{\boldsymbol\alpha=\hat{\boldsymbol\alpha}_n}\\
&\qquad\qquad \cdot \frac{\delta - F_{\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})}{F_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\{1-F_{nh,\hat{\boldsymbol\alpha}_n}(t-\hat{\boldsymbol\alpha}_n^T\boldsymbol{x})\}}\, dP_0(\boldsymbol{x},t,\delta)\\
&= \psi_{3\epsilon}'(\boldsymbol\alpha_0)(\hat{\boldsymbol\alpha}_n-\boldsymbol\alpha_0) + o_p\left(n^{-1/2}+\hat{\boldsymbol\alpha}_n-\boldsymbol\alpha_0\right). \qquad \text{(A.4.6)}
\end{aligned}
$$

The proof of (A.4.5) and (A.4.6) is similar to the proof of the corresponding steps given in the proof of Theorem 2.4.1 and omitted. □

Remark A.4.1. *It follows from the proof of Theorem 2.4.3 that*

$$\sqrt{n}\boldsymbol{I}_\epsilon(\boldsymbol{\alpha}_0)(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)$$
$$=n^{-1/2}\sum_{i=1}^n f_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)\{\mathbb{E}(\boldsymbol{X}_i|T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)-\boldsymbol{X}_i\}$$
$$\cdot\frac{\Delta_i-F_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)}{F_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)\{1-F_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)\}}1_{[\epsilon,1-\epsilon]}\left\{F_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)\right\}+o_p(1).$$

Therefore the result of Theorem 2.4.5 follows.

A.4.2 Estimation of the intercept

Proof of Theorem 2.5.1. We have

$$\begin{aligned}\hat\mu_n-\mu_0&=\int u\,dF_{nh,\hat{\boldsymbol{\alpha}}_n}(u)-\int u\,dF_0(u)=\int\left\{F_0(u)-F_{nh,\hat{\boldsymbol{\alpha}}_n}(u)\right\}du\\
&=\int\frac{F_0(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})-F_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}\,dG(\boldsymbol{x},t)\\
&=\int\frac{F_0(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})}{f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}\,dG(\boldsymbol{x},t)\\
&\quad+\int\frac{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})-F_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}{f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}\,dG(\boldsymbol{x},t).\end{aligned}\tag{A.4.7}$$

For the first term in the last expression we get:

$$\begin{aligned}&\int\frac{F_0(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})}{f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}\,dG(\boldsymbol{x},t)\\
&=\int\left\{F_0(u)-F_0(u+(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)^T\boldsymbol{x})\right\}f_{\boldsymbol{X}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(\boldsymbol{x}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}=u)\,du\,d\boldsymbol{x}\\
&\sim-\int(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)^T\boldsymbol{x}f_0(u)f_{\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X}}(\boldsymbol{x}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X}=u)\,du\,d\boldsymbol{x}\\
&\sim-(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)^T\left\{\int\mathbb{E}\{\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X}=u\}f_0(u)\,du\right\}.\end{aligned}$$

This term, multiplied with $\sqrt{n}$, is asymptotically normal, with expectation zero and variance

$$\sigma_1^2\stackrel{\text{def}}{=}a(\boldsymbol{\alpha}_0)^T\boldsymbol{I}_\epsilon(\boldsymbol{\alpha}_0)^{-1}\,a(\boldsymbol{\alpha}_0),$$

where $a(\boldsymbol{\alpha}_0)$ is the d-dimensional vector, defined by

$$a(\boldsymbol{\alpha}_0)=\int\mathbb{E}\{\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X}=u\}f_0(u)\,du.$$

For the second term in (A.4.7), we first note that

$$F_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x}) \\ = \frac{\int \{\delta - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\}K_h(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x} - u + \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{y})\, d\mathbb{P}_n(\boldsymbol{y}, u, \delta)}{g_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}. \tag{A.4.8}$$

We write (A.4.8) as the sum of the integral over dP_0 and the integral over $d(\mathbb{P}_n - P_0)$ and show that the contribution of the dP_0 integral, evaluated in (A.4.7) is negligible and that the contribution of the $d(\mathbb{P}_n - P_0)$ integral will yield an asymptotic normal distribution. We have:

$$\begin{aligned}
&\int \{\delta - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\}K_h(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x} - u + \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{y})\, dP_0(\boldsymbol{y}, u, \delta)\\
&= \int \{F_0(u - \boldsymbol{\alpha}_0^T\boldsymbol{y}) - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\}K_h(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x} - u + \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{y})\, dG(\boldsymbol{y}, u)\\
&= \int \{F_0(v + (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)y) - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\}\}K_h(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x} - v)\\
&\qquad\qquad \cdot f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(v) f_{\boldsymbol{X}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(\boldsymbol{y}|T - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X} = v)\, dv\, d\boldsymbol{y}\\
&= f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) \int \{F_0(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x} + (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)y) - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\}\\
&\qquad\qquad \cdot f_{\boldsymbol{X}|T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(\boldsymbol{y}|T - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X} = t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})\, d\boldsymbol{y} + O_p\left(h^2\right)\\
&= f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) f_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x}) \big(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\big)^T \mathbb{E}\{\boldsymbol{X} - \boldsymbol{x}|T - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X} = t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}\}\\
&\quad + O_p\left(h^2\right) + o_p\big(\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\|\big),
\end{aligned}$$

where $\|\boldsymbol{x}\|$ is the euclidean norm of the vector $\boldsymbol{x}$. Hence we get

$$\begin{aligned}
&\int \frac{\int \{\delta - F_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x})\}K_h(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x} - u + \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{y})\, dP_0(\boldsymbol{y}, u, \delta)}{g_{nh,\hat{\boldsymbol{\alpha}}_n}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}) f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}\, dG(\boldsymbol{x}, t)\\
&= (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)^T \int \frac{f_0(t - \boldsymbol{\alpha}_0^T\boldsymbol{x}) E\{\boldsymbol{X} - \boldsymbol{x}|T - \boldsymbol{\alpha}_0^T\boldsymbol{X} = t - \boldsymbol{\alpha}_0^T\boldsymbol{x}\}}{g_{nh}(t - \hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}\, dG(\boldsymbol{x}, t)\\
&\qquad + O_p\left(h^2\right) + o_p(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)\\
&= (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)^T \int f_0(v) \mathbb{E}\{\boldsymbol{X} - \boldsymbol{x}|T - \boldsymbol{\alpha}_0^T\boldsymbol{X} = v\} f_{\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X}}(\boldsymbol{x}|T - \boldsymbol{\alpha}_0^T\boldsymbol{X} = v)\, d\boldsymbol{x}\, dv\\
&\qquad + O_p\left(h^2\right) + o_p\big(\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\|\big)\\
&= O_p\left(h^2\right) + o_p\big(\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\|\big),
\end{aligned}$$

which is $o_p(n^{-1/2})$ if $h \ll n^{-1/4}$. Therefore, we use a bandwidth $h \asymp n^{-1/3}$ instead of $h \asymp n^{-1/5}$ in Theorem 2.5.1 to estimate the plug-in estimate $F_{nh,\hat{\boldsymbol{\alpha}}_n}$ of F_0. Since $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\| = O_p(n^{-1/2})$ by Theorem 2.4.3, we do not need to change the order of the bandwidth ($h \asymp n^{-1/5}$) for estimating $\boldsymbol{\alpha}_0$ in the estimation of the intercept.

Finally,

$$
\begin{aligned}
&\sqrt{n}\int\left(\frac{\int\{\delta-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\}K_h(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}-u+\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{y})\,d\big(\mathbb{P}_n-P_0\big)(\boldsymbol{y},u,\delta)}{g_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}\right)dG(\boldsymbol{x},t)\\
&=\sqrt{n}\int\left(\int\frac{\{\delta-F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\}K_h(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}-u+\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{y})}{g_{nh,\hat{\boldsymbol{\alpha}}_n}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})f_{T-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{X}}(t-\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})}\,dG(\boldsymbol{x},t)\right)d\big(\mathbb{P}_n-P_0\big)(\boldsymbol{y},u,\delta)\\
&=\sqrt{n}\int\frac{\{\delta-F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})\}}{f_{T-\boldsymbol{\alpha}_0^T\boldsymbol{X}}(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})}\,d\big(\mathbb{P}_n-P_0\big)(\boldsymbol{y},u,\delta)+O_p\left(h^2\right)+O_p\big(\|\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0\|\big),
\end{aligned}
$$

is asymptotically normal, with expectation zero and variance:

$$
\int\frac{F_0(v)\{1-F_0(v)\}}{f_{T-\boldsymbol{\alpha}_0^T\boldsymbol{X}}(v)}\,dv,
$$

if $h\ll n^{-1/4}$.

Both terms in the representation on the right of (A.4.7) are, apart from a negligible contribution, sums of independent variables with expectation zero. By Theorem 2.4.5 we have

$$
\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)=n^{-1/2}\boldsymbol{I}_\epsilon(\boldsymbol{\alpha}_0)^{-1}&\sum_{i=1}^n f_0(T_i-\boldsymbol{\alpha}_0X_i)\{\mathbb{E}(\boldsymbol{X}_i|T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)-\boldsymbol{X}_i\}\\
&\cdot\frac{\Delta_i-F_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)}{F_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)\{1-F_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)\}}1_{[\epsilon,1-\epsilon]}\left\{F_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)\right\}\\
&+o_p(1).
\end{aligned}
$$

and the second term of (A.4.7) has the representation:

$$
n^{-1/2}\sum_{i=1}^n\frac{\Delta_i-F_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)}{f_{T-\boldsymbol{\alpha}_0^T\boldsymbol{X}}(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)}.
$$

By the independence of the summands with indices $i\neq j$, the only contribution to the covariance of the two terms in the representation can come from summands with the same index. But,

$$
\begin{aligned}
&\mathbb{E}\Bigg\{\frac{f_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)\{\mathbb{E}(\boldsymbol{X}_i|T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)-\boldsymbol{X}_i\}\{\Delta_i-F_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)\}^2}{F_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)\{1-F_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)\}f_{T-\boldsymbol{\alpha}_0^T\boldsymbol{X}}(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)}\\
&\qquad\qquad\qquad\cdot 1_{[\epsilon,1-\epsilon]}\{F_0(T_i-\boldsymbol{\alpha}_0^T\boldsymbol{X}_i)\}\Bigg\}\\
&=\int_{F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})\in[\epsilon,1-\epsilon]} f_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})\{\mathbb{E}(\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X}=u-\boldsymbol{\alpha}_0^T\boldsymbol{y})-\boldsymbol{y}\}\\
&\qquad\cdot\frac{\{\delta-F_0(u-\boldsymbol{\alpha}_0^Ty)\}^2}{F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})\{1-F_0(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})\}f_{T-\boldsymbol{\alpha}_0^T\boldsymbol{X}}(u-\boldsymbol{\alpha}_0^T\boldsymbol{y})}dP_0(\boldsymbol{y},u,\delta)\\
&=\iint_{F_0(v)\in[\epsilon,1-\epsilon]} f_0(v)\{\mathbb{E}(\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X}=v)-\boldsymbol{y}\}\\
&\qquad\qquad\cdot\frac{F_0(v)\{1-F_0(v)\}}{F_0(v)\{1-F_0(v)\}}f_{\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X}}(\boldsymbol{y}|v)dvd\boldsymbol{y}\\
&=\int_{F_0(v)\in[\epsilon,1-\epsilon]}\left(\int\{\mathbb{E}(\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X}=v)-\boldsymbol{y}\}f_{\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X}}(\boldsymbol{y}|v)d\boldsymbol{y}\right)\\
&\qquad\qquad\frac{f_0(v)F_0(v)\{1-F_0(v)\}}{F_0(v)\{1-F_0(v)\}}dv\\
&=\mathbf{0}.
\end{aligned}
$$

So the covariance is zero and Theorem 2.5.1 follows.

□

Appendix B

Bootstrap procedures under current status data - Appendix

We give the proofs of Lemma 3.2.1 and Lemma 3.3.1 stated in Section 3.2 and Section 3.3 respectively. In Section B.1 we first proof Lemma 3.2.1 and derive the asymptotic normality result of the bootstrapped SMLE (3.2.2). The proof of Lemma 3.2.2 is given next. The behavior of the nonparametric bootstrap is given in Section B.2, where we also illustrate how the validity of the nonparametric bootstrap for generating the limiting distribution of the bootstrapped SMLE and the bootstrapped SSE is derived.

B.1 The smooth bootstrap

We introduce the notation

$$X_n = O_{\tilde{P}}(a_n),$$

to denote that for all $\varepsilon > 0$ and almost all sequences $(T_1, \Delta_1), (T_2, \Delta_2), \dots$, there exists a positive constant $K > 0$ such that

$$\tilde{P}_n \left\{ a_n^{-1} |X_n| \geq K \right\} < \varepsilon,$$

for all large n, where $\tilde{P}_n$ is the conditional probability measure of the (T_i, Δ_i^*), given $(T_1, \Delta_1), \dots, (T_n, \Delta_n)$. As a consequence of Lemma 3.2.1 we therefore get by an application of Markov's inequality that

$$\int_{t-h}^{t+h} \left\{ \hat{F}_n^*(x) - \tilde{F}_{nh}(x) \right\}^2 dx = O_{\tilde{P}} \left(h n^{-2/3} \right). \tag{B.1.1}$$

Recall that we denote the empirical measure of $(T_1, \Delta_1^*), \dots, (T_n, \Delta_n^*)$ by $\tilde{\mathbb{P}}_n$. In the proof of Lemma 3.2.1 we use the following (Dvoretsky-Kiefer-Wolfowitz-type) inequality from Banerjee et al. (2017).

Lemma B.1.1 (Lemma 8.1 of Banerjee et al. (2017)). *Let F be a distribution function on $\mathbb{R}$ with a density f supported on $[0,1]$ and bounded away from zero on $[0,1]$. Let $\mathbb{F}_n$ be the empirical distribution function associated with a sample of n observations from F and let $\mathbb{F}_n^{-1}$ be the corresponding empirical quantile function. With c a lower bound for f, we then have*

$$\mathbb{P}\left(\sup_{t\in[0,1]} |\mathbb{F}_n^{-1}(t) - F^{-1}(t)| > x\right) \le 4\exp(-2nc^2x^2),$$

for all n and $x > 0$.

B.1.1 Proof of Lemma 3.2.1

Just as in the proof of the corresponding Theorem 11.3 in Groeneboom and Jongbloed (2014), Doob's inequality and exponential centering play an important role in the proof. Moreover, we prove the equivalent statement

$$\tilde{P}_n\left\{\left|\tilde{U}_n(a) - \tilde{U}_0(a)\right| > x\right\} \le c_1 \exp\left\{-c_2 n x^3\right\}, \tag{B.1.2}$$

almost surely, for all large n, and constants $c_1, c_2 > 0$ and all $x \in (n^{-1/3}, M]$. To see that this is equivalent, first note that

$$\tilde{P}_n\left\{n^{1/3}\left|\tilde{U}_n(a) - \tilde{U}_0(a)\right| > x\right\} = \tilde{P}_n\left\{\left|\tilde{U}_n(a) - \tilde{U}_0(a)\right| > n^{-1/3}x\right\},$$

so, if (B.1.2) holds, we get that

$$\tilde{P}_n\left\{n^{1/3}\left|\tilde{U}_n(a) - \tilde{U}_0(a)\right| > x\right\} \le c_1 \exp\left\{-c_2 x^3\right\},$$

for all $x > 0$. Next note that for $x \in [0, n^{-1/3}]$

$$c_1 \exp\left\{-c_2 n x^3\right\} \ge c_1 \exp\left\{-c_2\right\} \ge 1,$$

if $c_1 \ge e^{c_2}$. So we can always adapt the constants in such a way that the inequality is satisfied for $x \in [0, n^{-1/3}]$. Furthermore, for $x \in [1, M]$, we can write:

$$c_1 \exp\left\{-c_2 n x^2\right\} \le c_1 \exp\left\{-(c_2/M) n x^3\right\}.$$

So for $x \in [1, M]$, we only need an inequality with x^2 in the exponent on the right-hand side, and can use Lemma B.1.1 to our advantage (see below). Finally, for $x > M$, the probability on the left-hand side of (B.1.2) is zero.

Let $\Lambda_n^* : [0,1] \to [0,1]$ be defined by $\Lambda_n^*(0) = 0$, and

$$\Lambda_n^*(i/n) = n^{-1} \sum_{j \le i} \Delta_j^*, \qquad 1 \le i \le n,$$

and by linear interpolation at other points of $[0,1]$. Furthermore, let λ_n^* be the left-continuous slope of the greatest convex minorant of Λ_n^*. Then:

$$\hat{F}_n^*(T_i) = \lambda_n^*(i/n) = \lambda_n^*(\mathbb{G}_n(T_i)),$$

where $\mathbb{G}_n$ is the empirical distribution function of the observations $T_1, \ldots, T_n$ and $\hat{F}_n^*$ is the MLE in the bootstrap sample.
We define analogously $\tilde{\lambda}_n = \tilde{F}_{nh} \circ G^{-1}$, and

$$\tilde{\Lambda}_n(t) = \int_0^t \tilde{\lambda}_n(u)\, du = \int_0^t \tilde{F}_{nh}\left(G^{-1}(u)\right)\, du, \qquad t \in [0,1].$$

Moreover, we define

$$W_n = \tilde{\lambda}_n^{-1}. \tag{B.1.3}$$

With these definitions we have that

$$\tilde{U}_0 = G^{-1} \circ \tilde{\lambda}_n^{-1} = G^{-1} \circ W_n. \tag{B.1.4}$$

By the model assumptions at the beginning of Section 3.1 for F_0 and G, and the almost sure convergence of $\tilde{F}_{nh}$ and its derivative to F_0 and f_0, respectively, uniformly on $[0,M]$ (using the suggested boundary correction near 0 and M), we may assume that there is a constant $c > 0$ such that $\tilde{\lambda}_n'(t) \ge c$ for all $t \in [0,1]$ and all large n, and that therefore, using a Taylor expansion, we get:

$$\tilde{\Lambda}_n(t) - \tilde{\Lambda}_n(W_n(a)) \ge \big(t - W_n(a)\big)a + \tfrac{1}{2}c\big(t - W_n(a)\big)^2, \tag{B.1.5}$$

for all $t, a \in [0,1]$.
We similarly define

$$W_n^*(a) = \text{argmin}_{u \in [0,1]}\{\Lambda_n^*(u) - au\},$$

where argmin denotes the smallest location of the minimum. Note that, analogously to (B.1.4), we have for $\tilde{U}_n$ as defined by (3.1.2) that

$$\tilde{U}_n = \mathbb{G}_n^{-1} \circ W_n^*. \tag{B.1.6}$$

By the transition of $\tilde{U}_0$ and $\tilde{U}_n$ to W_n and W_n^*, respectively, the range of $\tilde{U}_0$ and $\tilde{U}_n$ is changed from $[0,M]$ to $[0,1]$. We now prove:

$$\tilde{P}_n\left\{\left|W_n^*(a) - W_n(a)\right| > x\right\} \le c_1 \exp\left\{-c_2 n x^3\right\}, \tag{B.1.7}$$

almost surely, for all large n, and constants $c_1, c_2 > 0$ and all $x \in (n^{-1/3}, 1]$. Note that the probability on the left-hand side of (B.1.7) is zero if $x > 1$.
Define:

$$\varepsilon_i^* = \Delta_i^* - \tilde{F}_{nh}(T_i), \qquad 1 \le i \le n.$$

Then, we get that

$$\begin{aligned}\Lambda_n^*(i/n) &= n^{-1}\sum_{j\le i}\varepsilon_j^* + n^{-1}\sum_{j\le i}\tilde{F}_{nh}\left(\mathbb{G}_n^{-1}(j/n)\right)\\ &= n^{-1}\sum_{j\le i}\varepsilon_j^* + \int_0^{i/n}\tilde{F}_{nh}\left(\mathbb{G}_n^{-1}(u)\right)du, \qquad 1 \le i \le n,\end{aligned}$$

using the piecewise constancy of $\mathbb{G}_n^{-1}$. This gives:

$$\begin{aligned}&\tilde{P}_n\left\{\left|W_n^*(a) - W_n(a)\right| > x\right\}\\ &\le \tilde{P}_n\left\{\min_{i:\ |W_n(a)-i/n|>x}\{\Lambda_n^*(i/n) - a\,i/n\} \le \Lambda_n^*(W_n(a)) - aW_n(a)\right\}\\ &\le \tilde{P}_n\left\{\min_{i:\ |W_n(a)-i/n|>x}\left\{D_n^*(i/n) - D_n^*\left(W_n(a)\right) + \tfrac{1}{2}c\left(in^{-1} - W_n(a)\right)^2\right\} \le 0\right\},\end{aligned}$$

where D_n^* is defined by $D_n^* = \Lambda_n^* - \tilde{\Lambda}_n$ and where we use (B.1.5) in the last step. Define:

$$B_n^*(t) = D_n^*(t) - \int_0^t\left\{\tilde{F}_{nh}\left(\mathbb{G}_n^{-1}(u)\right) - \tilde{F}_{nh}\left(G^{-1}(u)\right)\right\}du.$$

Then we get that

$$B_n^*(i/n) = n^{-1}\sum_{j\le i}\varepsilon_i^*.$$

Moreover, the event $\{\left|W_n^*(a) - W_n(a)\right| > x\}$ is contained in the union of the events

$$E_{n1} =$$
$$\left\{\sup_{u:\ |W_n(a)-u|>x}\left\{\int_{W_n(a)}^u\left\{\tilde{F}_{nh}\left(G^{-1}(t)\right) - \tilde{F}_{nh}\left(\mathbb{G}_n^{-1}(t)\right)\right\}dt - \frac{c}{4}(u - W_n(a))^2\right\} \ge 0\right\}$$

and

$$E_{n2} = \left\{\sup_{i:\ |W_n(a)-i/n|>x}\{B_n^*(W_n(a)) - B_n^*(i/n) - \frac{c}{4}(in^{-1} - W_n(a))^2\} \ge 0\right\}.$$

We have, by the mean value theorem and the bounded differentiability of $\tilde{F}_{nh}$,

$$\begin{aligned}&\left|\int_{W_n(a)}^u\left\{\tilde{F}_{nh}\left(G^{-1}(t)\right) - \tilde{F}_{nh}\left(\mathbb{G}_n^{-1}(t)\right)\right\}dt\right|\\ &\qquad\le c'\left|u - W_n(a)\right|\sup_{t\in[0,1]}\left|\mathbb{G}_n^{-1}(t) - G^{-1}(t)\right|,\end{aligned}$$

for a constant $c' > 0$. Hence we get from Lemma B.1.1 in the original space that

$$P_n(E_{n1}) \le P_n\left\{\sup_{t\in[0,1]}\left|\mathbb{G}_n^{-1}(t) - G^{-1}(t)\right| \ge \frac{cx}{4c'}\right\} \le 4\exp\left\{-Knc^2x^2\right\}$$
$$\le 4\exp\left\{-Kc^2n^{1/3}\right\}, \qquad \text{(B.1.8)}$$

for some $K > 0$ and $x \in (n^{-1/3}, M]$. This means that we may assume that, almost surely, the complement of E_{n1} is satisfied for all large n and all $x \in (n^{-1/3}, 1]$. So we now turn to $\tilde{P}_n(E_{n2})$. We have that

$$\tilde{P}_n(E_{n2})$$
$$\le \sum_{k\ge1}\tilde{P}_n\left(\sup_{i:\ |W_n(a)-i/n|\in(kx,(k+1)x]}\left\{B_n^*\left(W_n(a)\right) - B_n^*(i/n) - \frac{c}{4}\left(i/n - W_n(a)\right)^2\right\} \ge 0\right)$$
$$\le \sum_{k\ge1}\tilde{P}_n\left(\sup_{i:\ |W_n(a)-i/n|\le(k+1)x}\left\{B_n^*\left(W_n(a)\right) - B_n^*(i/n)\right\} \ge \frac{c}{4}k^2x^2\right).$$

Using the piecewise linearity of B_n^*, we get that

$$B_n^*\left(W_n(a)\right) = B_n^*\left(\frac{\lfloor nW_n(a)\rfloor}{n}\right) + \left(W_n(a) - \frac{\lfloor nW_n(a)\rfloor}{n}\right)\varepsilon^*_{\lfloor nW_n(a)\rfloor+1},$$

where $\lfloor nW_n(a)\rfloor$ denotes the integer part ("floor") of $nW_n(a)$. Hence,

$$\tilde{P}_n(E_{n2}) \le \sum_{k\ge1}\tilde{P}_n\left(\left(W_n(a) - \frac{\lfloor nW_n(a)\rfloor}{n}\right)\varepsilon^*_{\lfloor nW_n(a)\rfloor+1} \ge \frac{c}{8}k^2x^2\right)$$
$$+\sum_{k\ge1}\tilde{P}_n\left(\sup_{i:\ |W_n(a)-i/n|\le(k+1)x}\left\{\sum_{j\le nW_n(a)}\varepsilon_j^* - \sum_{j\le i}\varepsilon_j^*\right\} \ge \frac{nc}{8}k^2x^2\right). \qquad \text{(B.1.9)}$$

The Markov inequality implies that for all $\theta > 0$, $k \ge 1$, $a \in [0,1]$ and $x \in (n^{-1/3}, 1]$,

$$\tilde{P}_n\left\{\left(W_n(a) - \frac{\lfloor nW_n(a)\rfloor}{n}\right)\varepsilon^*_{\lfloor nW_n(a)\rfloor+1} \ge \frac{c}{8}k^2x^2\right\}$$
$$\le \exp\left\{-\frac{\theta c}{8}k^2x^2\right\}\tilde{E}_n\exp\left\{\theta\left(W_n(a) - \frac{\lfloor nW_n(a)\rfloor}{n}\right)\varepsilon^*_{\lfloor nW_n(a)\rfloor+1}\right\},$$

where $\tilde{E}_n$ denotes the expectation under $\tilde{P}_n$. Since $\varepsilon_i^* \in [-1, 1]$ for all i, we have $\exp(\alpha\varepsilon_i^*) \le K\exp(\alpha^2)$ for all $\alpha \in \mathbb{R}$ and $K \ge \exp(1)$ and therefore, with $\theta = ck^2x^2n^2/16$, we obtain

$$\tilde{P}_n\left(\left(W_n(a) - \frac{\lfloor nW_n(a)\rfloor}{n}\right)\varepsilon^*_{\lfloor nW_n(a)\rfloor+1} \ge \frac{c}{8}k^2x^2\right) \le K\exp\left(-\frac{\theta c}{8}k^2x^2 + \frac{\theta^2}{n^2}\right)$$
$$\le K\exp\left(-\frac{c^2k^4x^4n^2}{256}\right).$$

Using that $k^4 \geq k$ for all $k \geq 1$ and $nx \geq 1$ for all $x \in (n^{-1/3}, 1)$, we conclude that for all $a \in [0,1]$ and $x \in (n^{-1/3}, 1)$

$$\begin{aligned}
&\sum_{k \geq 1} \tilde{P}_n\left(\left(W_n(a) - \frac{\lfloor nW_n(a)\rfloor}{n}\right)\varepsilon^*_{\lfloor nW_n(a)\rfloor+1} \geq \frac{c}{8}k^2x^2\right) \\
&\leq K\sum_{k\geq 1}\exp\left(-\frac{c^2kx^3n}{256}\right) \leq K\exp\left(-\frac{c^2x^3n}{256}\right)\sum_{k\geq 0}\exp\left(-\frac{c^2kx^3n}{256}\right) \\
&\leq K'\exp(-K_2nx^3), \qquad \text{(B.1.10)}
\end{aligned}$$

with any finite K' that satisfies $K' \geq K\sum_{k\geq 0}\exp\left(-c^2k/256\right)$ and $K_2 \leq c^2/256$. This takes care of the first term on the right of (B.1.9).

We now consider the second term on the right of (B.1.9). Just as in the proof of Theorem 11.3 in Groeneboom and Jongbloed (2014), we use Doob's submartingale inequality, this time conditionally on $(T_1, \Delta_1), \ldots, (T_n, \Delta_n)$. This gives

$$\begin{aligned}
&\tilde{P}_n\left(\sup_{i:\ |W_n(a)-i/n|\leq(k+1)x}\left\{\sum_{j\leq nW_n(a)}\varepsilon^*_j - \sum_{j\leq i}\varepsilon^*_j\right\} \geq \frac{nc}{8}k^2x^2\right) \\
&\leq 2\exp\left(-\frac{\theta nc}{8}k^2x^2\right)\sup_{i:\ |W_n(a)-i/n|\leq(k+1)x}\tilde{E}_n\left[\exp\left(\theta\left(\sum_{j\leq nW_n(a)}\varepsilon^*_j - \sum_{j\leq i}\varepsilon^*_j\right)\right)\right]. \qquad \text{(B.1.11)}
\end{aligned}$$

Suppose $i/n < W_n(a)$. Then we get that

$$\begin{aligned}
&\log\tilde{E}_n\left[\exp\left(\theta\left(\sum_{j\leq nW_n(a)}\varepsilon^*_j - \sum_{j\leq i}\varepsilon^*_j\right)\right)\right] = \log\tilde{E}_n\left[\exp\left(\theta\left(\sum_{i<j\leq nW_n(a)}\varepsilon^*_j\right)\right)\right] \\
&= \sum_{i<j\leq nW_n(a)}\log\left\{\exp\left\{\theta\{1-\tilde{F}_{nh}(T_j)\}\right\}\tilde{F}_{nh}(T_j) + \exp\left\{-\theta\tilde{F}_{nh}(T_j)\right\}\{1-\tilde{F}_{nh}(T_j)\}\right\} \\
&= n\int_{i/n}^{W_n(a)}\log\left\{\exp\left\{\theta\{1-\tilde{F}_{nh}(\mathbb{G}_n^{-1}(t))\}\right\}\tilde{F}_{nh}(\mathbb{G}_n^{-1}(t))\right. \\
&\qquad\qquad \left. + \exp\left\{-\theta\tilde{F}_{nh}(\mathbb{G}_n^{-1}(t))\right\}\left\{1-\tilde{F}_{nh}(\mathbb{G}_n^{-1}(t))\right\}\right\}dt.
\end{aligned}$$

Since $\log(1+x) \le x$, this is bounded above by

$$
\begin{aligned}
& n \int_{i/n}^{W_n(a)} \Big\{ \exp\left\{\theta\{1-\tilde F_{nh}(\mathbb{G}_n^{-1}(t))\}\right\} \tilde F_{nh}(\mathbb{G}_n^{-1}(t)) \\
&\qquad\qquad + \exp\left\{-\theta \tilde F_{nh}(\mathbb{G}_n^{-1}(t))\right\} \left\{1-\tilde F_{nh}(\mathbb{G}_n^{-1}(t))\right\} - 1 \Big\}\, dt \\
&\le n \int_{W_n(a)-(k+1)x}^{W_n(a)} \Big\{ \exp\left\{\theta\{1-\tilde F_{nh}(\mathbb{G}_n^{-1}(t))\}\right\} \tilde F_{nh}(\mathbb{G}_n^{-1}(t)) \\
&\qquad\qquad + \exp\left\{-\theta \tilde F_{nh}(\mathbb{G}_n^{-1}(t))\right\} \left\{1-\tilde F_{nh}(\mathbb{G}_n^{-1}(t))\right\} - 1 \Big\}\, dt, \\
&= n \int_{W_n(a)-(k+1)x}^{W_n(a)} \Big\{ \sum_{i=2}^{\infty} \frac{\theta^i}{i!} \{1-\tilde F_{nh}(\mathbb{G}_n^{-1}(t))\}^i \tilde F_{nh}(\mathbb{G}_n^{-1}(t)) \\
&\qquad\qquad + \sum_{i=2}^{\infty} \frac{\theta^i}{i!} (-1)^i \tilde F_{nh}(\mathbb{G}_n^{-1}(t))^i \left\{1-\tilde F_{nh}(\mathbb{G}_n^{-1}(t))\right\} \Big\}\, dt, \\
&= n \sum_{i=2}^{\infty} \frac{\theta^i}{i!} \int_{W_n(a)-(k+1)x}^{W_n(a)} \Big\{ \{1-\tilde F_{nh}(\mathbb{G}_n^{-1}(t))\}^i \tilde F_{nh}(\mathbb{G}_n^{-1}(t)) \\
&\qquad\qquad + (-1)^i \tilde F_{nh}(\mathbb{G}_n^{-1}(t))^i \left\{1-\tilde F_{nh}(\mathbb{G}_n^{-1}(t))\right\} \Big\}\, dt,
\end{aligned}
$$

if $i/n < W_n(a)$ and $|W_n(a) - i/n| \le (k+1)x$. Since $W_n(a) \in [0,1]$, the integrand,

$$
\{1-\tilde F_{nh}(\mathbb{G}_n^{-1}(t))\}^i \tilde F_{nh}(\mathbb{G}_n^{-1}(t)) + (-1)^i \tilde F_{nh}(\mathbb{G}_n^{-1}(t))^i \left\{1-\tilde F_{nh}(\mathbb{G}_n^{-1}(t))\right\},
$$

is bounded by $1/2$, we get,

$$
\begin{aligned}
& \tilde P_n \left(\sup_{i:\ |W_n(a)-i/n| \le (k+1)x} \left\{ \sum_{j \le nW_n(a)} \varepsilon_j^* - \sum_{j \le i} \varepsilon_j^* \right\} \ge \frac{nc}{8} k^2 x^2 \right) \\
&\le 2 \exp\left(-\frac{\theta nc}{8} k^2 x^2\right) \sup_{i:\ |W_n(a)-i/n| \le (k+1)x} \tilde E_n \left[\exp\left(\theta \left(\sum_{j \le nW_n(a)} \varepsilon_j^* - \sum_{j \le i} \varepsilon_j^* \right)\right)\right] \\
&\le 2 \exp\left(-\frac{\theta n c k^2 x^2}{8} + \frac{n(k+1)x}{2} \sum_{i=2}^{\infty} \frac{\theta^i}{i!} \right),
\end{aligned}
$$

for all $x \in (n^{-1/3}, 1), k \ge 1, a \in [0,1]$ and $\theta > 0$. Therefore, with $\theta = \log(1 + \frac{ck^2 x}{4(k+1)})$, we arrive at,

$$\tilde{P}_n\left(\sup_{i:\ |W_n(a)-i/n|\le(k+1)x}\left\{\sum_{j\le nW_n(a)}\varepsilon_j^*-\sum_{j\le i}\varepsilon_j^*\right\}\ge\frac{nc}{8}k^2x^2\right)$$
$$\le 2\exp\left(\frac{n(k+1)x}{2}\left\{\frac{ck^2x}{4(k+1)}-\left(1+\frac{ck^2x}{4(k+1)}\right)\log\left(1+\frac{ck^2x}{4(k+1)}\right)\right\}\right).$$

Following Pollard (1984), in his discussion of Bennett's inequality on p. 192, we introduce the function B, defined by $B(0)=1/2$ and

$$B(u)=u^{-2}\{(1+u)\log(1+u)-u\}.$$

Making the change of variables $u_k=ck^2x/(4(k+1))$, we can write

$$\tilde{P}_n\left(\sup_{i:\ |W_n(a)-i/n|\le(k+1)x}\left\{\sum_{j\le nW_n(a)}\varepsilon_j^*-\sum_{j\le i}\varepsilon_j^*\right\}\ge\frac{nc}{8}k^2x^2\right)$$
$$\le 2\exp\left(-\frac{nc^2k^4x^3}{32(k+1)}B(u_k)\right).$$

Since u_k varies over a finite interval $[0,M']$ and therefore $B(u_k)$ stays away from zero on $[0,M']$, we find that

$$\sum_{k\ge1}\tilde{P}_n\left(\sup_{i:\ |W_n(a)-i/n|\le(k+1)x}\left\{\sum_{j\le nW_n(a)}\varepsilon_j^*-\sum_{j\le i}\varepsilon_j^*\right\}\ge\frac{nc}{8}k^2x^2\right)$$
$$\le K_1\sum_{k\ge1}\exp\left(-\frac{nc^2k^4x^3}{32(k+1)}\right)\le K_1\exp\left(-\frac{nc^2x^3}{64}\right)\sum_{k\ge0}\exp\left(-\frac{c^2k^4}{32(k+1)}\right)$$
$$\le K_2\exp\left(-K_3nx^3\right).$$

for appropriate K_1, K_2 and K_3. Combining this with (B.1.8) and (B.1.10), it follows that

$$\tilde{P}_n\left\{\left|W_n^*(a)-W_n(a)\right|>x\right\}\le c_1\exp\{-nc_2x^3\},$$

for all large n, almost surely along $(T_1,\Delta_1),\ldots$ for constants $c_1,c_2>0$ and $x\in(n^{-1/3},1]$. We now prove that (B.1.2) also follows by considering the transition of W_n and W_n^* to $\tilde{U}_0$ and $\tilde{U}_n$. By (B.1.4) and (B.1.6) we get that

$$\tilde{U}_n(a)-\tilde{U}_0(a)=\mathbb{G}_n^{-1}\circ W_n^*(a)-G^{-1}\circ W_n(a),$$

and hence

$$\left|\tilde{U}_n(a)-\tilde{U}_0(a)\right|\le\sup_{t\in[0,1]}\left|\mathbb{G}_n^{-1}(t)-G^{-1}(t)\right|+k_1\left|W_n^*(a)-W_n(a)\right|,$$

where

$$k_1 = 1/\inf_{x\in[0,M]} g(x).$$

From Lemma B.1.1 we get in the original space that

$$P_n\left\{\sup_{t\in[0,1]}\left|\mathbb{G}_n^{-1}(t) - G^{-1}(t)\right| \geq x/2\right\} \leq 4\exp\left\{-Kn^{1/3}\right\},$$

for some $K > 0$ and $x \in (n^{-1/3}, M]$. So we may assume that, almost surely, $\left|\mathbb{G}_n^{-1}(t) - G^{-1}(t)\right| < x/2$, for all large n and all $x \in (n^{-1/3}, 1]$. By the foregoing proof, we also have that

$$\tilde{P}_n\left\{k_1\left|W_n^*(a) - W_n(a)\right| \geq x/2\right\} \leq c_1\exp\left\{-c_2nx^3/\left(8k_1^3\right)\right\}.$$

This proves the result.

B.1.2 Asymptotic normality of the bootstrapped SMLE

In this section, we prove (3.2.2). Define the functions

$$\psi_{t,h}(u) = \frac{K_h(t-u)}{g(u)}, \tag{B.1.12}$$

and

$$\bar{\psi}^*_{t,h}(u) = \begin{cases} \psi_{t,h}(\tau_i), & \text{if } \tilde{F}_{nh}(u) > \hat{F}_n^*(\tau_i),\ u \in [\tau_i, \tau_{i+1}),\\ \psi_{t,h}s), & \text{if } \tilde{F}_{nh}(u) = \hat{F}_n^*(s),\ \text{for some } s \in [\tau_i, \tau_{i+1}),\\ \psi_{t,h}(\tau_{i+1}), & \text{if } \tilde{F}_{nh}(u) < \hat{F}_n^*(\tau_i),\ u \in [\tau_i, \tau_{i+1}), \end{cases}$$

where the τ_i are the points of jump of $\hat{F}_n^*$. By the convex minorant interpretation of $\hat{F}_n^*$ we have that

$$\int \bar{\psi}^*_{t,h}(u)\left\{\delta^* - \hat{F}_n^*(u)\right\} d\tilde{\mathbb{P}}_n(u,\delta^*) = 0.$$

This implies that

$$\begin{aligned}
0 &= \int \bar{\psi}^*_{t,h}(u)\left\{\delta^* - \hat{F}_n^*(u)\right\} d\tilde{\mathbb{P}}_n(u,\delta^*)\\
&= \int \psi_{t,h}(u)\left\{\delta^* - \hat{F}_n^*(u)\right\} d\tilde{\mathbb{P}}_n(u,\delta^*) + \int \left\{\bar{\psi}^*_{t,h}(u) - \psi_{t,h}(u)\right\}\left\{\delta^* - \hat{F}_n^*(u)\right\} d\tilde{\mathbb{P}}_n(u,\delta^*)\\
&= \int \psi_{t,h}(u)\left\{\delta^* - \tilde{F}_{nh}^*(u)\right\} d(\tilde{\mathbb{P}}_n - \tilde{P}_n)(u,\delta^*)\\
&\quad + \int \psi_{t,h}(u)\left\{\tilde{F}_{nh}^*(u) - \hat{F}_n^*(u)\right\} d\tilde{\mathbb{P}}_n(u,\delta^*)\\
&\quad + \int \left\{\bar{\psi}^*_{t,h}(u) - \psi_{t,h}(u)\right\}\left\{\delta^* - \hat{F}_n^*(u)\right\} d\tilde{\mathbb{P}}_n(u,\delta^*),
\end{aligned}$$

where we write $d(\tilde{\mathbb{P}}_n - \tilde{P}_n)$ instead of $d\tilde{\mathbb{P}}_n$ in the last equality as a result of the fact that

$$\tilde{P}_n\left(\Delta_i^* = 1\right) = \tilde{F}_{nh}(T_i) \quad 1 \le i \le n. \tag{B.1.13}$$

Using integrating by parts we have

$$\tilde{F}_{nh}^*(t) - \int \mathbb{K}_h(t-u)\, d\tilde{F}_{nh}(u) = \int \psi_{t,h}(u)\, \left\{\hat{F}_n^*(u) - \tilde{F}_{nh}(u)\right\} dG(u).$$

So we find that

$$\begin{aligned}
&\tilde{F}_{nh}^*(t) - \int \mathbb{K}_h(t-u)\, d\tilde{F}_{nh}(u) \\
&= \int \bar{\psi}_{t,h}^*(u) \left\{\delta^* - \hat{F}_n^*(u)\right\} d\tilde{\mathbb{P}}_n(u,\delta^*) - \int \psi_{t,h}(u)\, \left\{\tilde{F}_{nh}(u) - \hat{F}_n^*(u)\right\} dG(u) \\
&= \int \psi_{t,h}(u) \left\{\delta^* - \tilde{F}_{nh}(u)\right\} d(\tilde{\mathbb{P}}_n - \tilde{P}_n)(u,\delta^*) \\
&\quad + \int \psi_{t,h}(u) \left\{\tilde{F}_{nh}(u) - \hat{F}_n^*(u)\right\} d(\mathbb{G}_n - G)(u,\delta^*) \\
&\quad + \int \left\{\bar{\psi}_{t,h}^*(u) - \psi_{t,h}(u)\right\} \left\{\delta^* - \hat{F}_n^*(u)\right\} d\tilde{\mathbb{P}}_n(u,\delta^*) \\
&= A_I + A_{II} + A_{III}.
\end{aligned}$$

To study the asymptotic distribution of

$$n^{2/5} \left\{\tilde{F}_{nh}^*(t) - \int \mathbb{K}_h(t-u)\, d\tilde{F}_{nh}(u)\right\},$$

we therefore have to analyze the three terms A_I, A_{II} and A_{III}. We start with A_I and prove that

$$n^{2/5} \int \psi_{t,h}(u) \left\{\delta^* - \tilde{F}_{nh}(u)\right\} d(\tilde{\mathbb{P}}_n - \tilde{P}_n)(u,\delta^*) \xrightarrow{\mathcal{D}} N\left(0, \sigma^2\right), \tag{B.1.14}$$

where σ^2 is defined in (3.2.3). Define

$$Z_{nh,i} = n^{-3/5} \psi_{t,h}(T_i) \left\{\Delta_i^* - \tilde{F}_{nh}(T_i)\right\}.$$

The left hand side of (B.1.14) can then be expressed as $\sum_{i=1}^n Z_{nh,i}$. Conditionally on $(T_1, X_1), \ldots, (T_n, X_n)$, $Z_{nh,i}$ has mean zero and variance

$$\sigma_{nh,i}^2 = n^{-6/5} \psi_{t,h}^2(T_i) \tilde{F}_{nh}(T_i) \left\{1 - \tilde{F}_{nh}(T_i)\right\}.$$

Therefore, along almost all sequences $(T_1, \Delta_1), (T_2, \Delta_2) \ldots$,

$$\begin{aligned}\sum_{i=1}^{n} \sigma_{nh,i}^2 &= n^{-1/5} \int \psi_{t,h}^2(u) \tilde{F}_{nh}(u) \left\{1 - \tilde{F}_{nh}(u)\right\} d\mathbb{G}_n(u) \\ &= n^{-1/5} \int \psi_{t,h}^2(u) \tilde{F}_{nh}(u) \left\{1 - \tilde{F}_{nh}(u)\right\} dG(u) + o(1) \\ &= \int_{-1}^{1} K^2(u) \tilde{F}_{nh}(t+hu) \left\{1 - \tilde{F}_{nh}(t+hu)\right\} g(t+hu) du + o(1) \\ &\to \frac{F_0(t)\{1 - F_0(t)\}}{cg(t)} \int K^2(u) du = \sigma^2,\end{aligned}$$

where we use the a.s. convergence of $F_{nh}(t) \to F_0(t)$ in the last line. By the Lindeberg-Feller CLT, we have,

$$\sum_{i=1}^{n} Z_{nh,i} \xrightarrow{\mathcal{D}} N\left(0, \sigma^2\right).$$

This proves (B.1.14).

We next consider A_{II}. From the fact that the integrand is the product of h^{-1} times the fixed bounded continuous function $u \mapsto K((t-u)/h)/g(u)$ and the class of functions of bounded variation $\hat{F}_n^* - \tilde{F}_{nh}$ which have entropy with bracketing of order ε^{-1} for the L_2-distance and are of order $O_{\tilde{P}}(n^{-1/3})$ for the L_2-distance, again conditionally on $\omega = (T_1, \Delta_1), (T_2, \Delta_2), \ldots$, it follows that A_{II} is of order $O_{\tilde{P}}(h^{-1} n^{-2/3})$. As a consequence, we have for $h \asymp n^{-1/5}$,

$$A_{II} = \int \psi_{t,h}(u) \left\{\tilde{F}_{nh}(u) - \hat{F}_n^*(u)\right\} d(\mathbb{G}_n - G)(u) = o_{\tilde{P}}(n^{-2/5}) \tag{B.1.15}$$

We finally study the term A_{III}. Using similar arguments as in the proof of Lemma A.4 in Groeneboom et al. (2010), there exists a positive constant C such that

$$\left|\bar{\psi}_{t,h}^*(u) - \psi_{t,h}(u)\right| \le Ch^{-2} \left|\hat{F}_n^*(u) - \tilde{F}_{nh}(u)\right|, \tag{B.1.16}$$

for all u such that $\tilde{f}_{nh} = \tilde{F}_{nh}'$ is positive and continuous in a neighborhood around u. By (B.1.13), we can write,

$$\begin{aligned}A_{III} &= \int \left\{\bar{\psi}_{t,h}^*(u) - \psi_{t,h}(u)\right\} \left\{\delta^* - \tilde{F}_{nh}(u)\right\} d(\tilde{\mathbb{P}}_n - \tilde{P}_n)(u, \delta^*) \\ &\quad + \int \left\{\bar{\psi}_{t,h}^*(u) - \psi_{t,h}(u)\right\} \left\{\tilde{F}_{nh}(u) - \hat{F}_n^*(u)\right\} d\mathbb{G}_n(u).\end{aligned} \tag{B.1.17}$$

It is clear that

$$\begin{aligned}&\int \left\{\bar{\psi}_{t,h}^*(u) - \psi_{t,h}(u)\right\} \left\{\delta^* - \tilde{F}_{nh}(u)\right\} d(\tilde{\mathbb{P}}_n - \tilde{P}_n)(u, \delta^*) \\ &\quad = o_{\tilde{P}}\left(\int \psi_{t,h}(u) \left\{\delta^* - \tilde{F}_{nh}(u)\right\} d(\tilde{\mathbb{P}}_n - \tilde{P}_n)(u, \delta^*)\right),\end{aligned}$$

which is $o_{\tilde{P}}(n^{-2/5})$ by (B.1.14). For the second term on the right-hand side of (B.1.17) we get by (B.1.1) and (B.1.16) that

$$\left|\int \left\{\bar{\psi}^*_{t,h}(u) - \psi_{t,h}(u)\right\} \left\{\tilde{F}_{nh}(u) - \hat{F}^*_n(u)\right\} d\mathbb{G}_n(u)\right|$$
$$\le Ch^{-2} \int_{t-h}^{t+h} \left\{\tilde{F}_{nh}(u) - \hat{F}^*_n(u)\right\}^2 d\mathbb{G}_n(u) = O_{\tilde{P}}\left(h^{-1}n^{-2/3}\right) = O_{\tilde{P}}\left(n^{-7/15}\right). \tag{B.1.18}$$

The asymptotic normality of the bootstrapped SMLE given in (3.2.2) now follows by (B.1.14),(B.1.15) and (B.1.18).

B.1.3 Proof of Lemma 3.2.2

We have:

$$\int \left\{\mathbb{K}_h(t-u) + \mathbb{K}_h(t+u) - \mathbb{K}_h(2M-t-u)\right\} d\tilde{F}^{(bc)}_{nh}(u)$$
$$= \int_{u=0}^{M} \left\{\mathbb{K}_h(t-u) + \mathbb{K}_h(t+u) - \mathbb{K}_h(2M-t-u)\right\} \tilde{f}^{(bc)}_{nh}(u)\,du.$$

If $t \in [h, M-h]$ we get, noting that $\mathbb{K}_h(t+u) = \mathbb{K}_h(2M-t-u) = 1$, if $t \in [h, M-h]$,

$$\int_{u=0}^{M} \left\{\mathbb{K}_h(t-u) + \mathbb{K}_h(t+u) - \mathbb{K}_h(2M-t-u)\right\} \tilde{f}^{(bc)}_{nh}(u)\,du$$
$$= \int_{u=0}^{M} \mathbb{K}_h(t-u)\tilde{f}^{(bc)}_{nh}(u)\,du$$
$$= \int_{u=0}^{M} \mathbb{K}_h(t-u) \int \left\{K_h(u-v) + K_h(u+v) + K_h(2M-u-v)\right\} d\hat{F}_n(v)\,du$$
$$= \int \left\{\int_{u=0}^{M} \mathbb{K}_h(t-u)\left\{K_h(u-v) + K_h(u+v) + K_h(2M-u-v)\right\} du\right\} d\hat{F}_n(v)$$
$$= \int \left\{\widetilde{\mathbb{K}}_h(t-v) + \widetilde{\mathbb{K}}_h(t+v) - \widetilde{\mathbb{K}}_h(2M-t-v)\right\} d\hat{F}_n(v)$$

The last transition follows from integration by parts and the symmetry of the kernel K

$$
\begin{aligned}
&\int_{u=0}^{M} \mathbb{K}_h(t-u)\left\{K_h(u-v)+K_h(u+v)+K_h(2M-u-v)\right\}\,du\\
&=\left[\mathbb{K}_h(t-u)\left\{\mathbb{K}_h(u-v)+\mathbb{K}_h(u+v)-\mathbb{K}_h(2M-u-v)\right\}\right]_{u=0}^{M}\\
&\qquad\qquad+\int K_h(t-u)\left\{\mathbb{K}_h(u-v)+\mathbb{K}_h(u+v)-\mathbb{K}_h(2M-u-v)\right\}\,du\\
&=\int K_h(t-u)\left\{\mathbb{K}_h(u-v)+\mathbb{K}_h(u+v)-\mathbb{K}_h(2M-u-v)\right\}\,du\\
&=\int \left\{\mathbb{K}_h(t-v-hw)+\mathbb{K}_h(t+v-hw)-\mathbb{K}_h(2M-t-v-hw)\right\}K(w)\,dw\\
&=\widetilde{\mathbb{K}}_h(t-v)+\widetilde{\mathbb{K}}_h(t+v)-\widetilde{\mathbb{K}}_h(2M-t-v).
\end{aligned}
$$

if $t\in[h,M-h]$.

We likewise get, if $t\in[0,h]$,

$$
\begin{aligned}
&\int_{u=0}^{M}\left\{\mathbb{K}_h(t-u)+\mathbb{K}_h(t+u)-\mathbb{K}_h(2M-t-u)\right\}\tilde f_{nh}^{(bc)}(u)\,du\\
&=\int_{u=0}^{M}\left\{\mathbb{K}_h(t-u)+\mathbb{K}_h(t+u)-1\right\}\tilde f_{nh}^{(bc)}(u)\,du\\
&=\int_{u=0}^{M}\left\{\mathbb{K}_h(t-u)+\mathbb{K}_h(t+u)-1\right\}\\
&\qquad\qquad\cdot\int\left\{K_h(u-v)+K_h(u+v)+K_h(2M-u-v)\right\}\,d\hat F_n(v)\,du\\
&=\int\left\{\widetilde{\mathbb{K}}_h(t-v)+\widetilde{\mathbb{K}}_h(t+v)-\widetilde{\mathbb{K}}_h(2M-t-v)\right\}\,d\hat F_n(v).
\end{aligned}
$$

In the last transition we use integration by parts again:

$$
\begin{aligned}
&\int_{u=0}^{M}\left\{\mathbb{K}_h(t-u)+\mathbb{K}_h(t+u)-1\right\}\\
&\qquad\qquad\cdot\left\{K_h(u-v)+K_h(u+v)+K_h(2M-u-v)\right\}\,du\\
&=\left[\left\{\mathbb{K}_h(t-u)+\mathbb{K}_h(t+u)-1\right\}\left\{\mathbb{K}_h(u-v)+\mathbb{K}_h(u+v)-\mathbb{K}_h(2M-u-v)\right\}\right]_{u=0}^{M}\\
&\quad+\int_{u=0}^{M}\left\{K_h(t-u)-K_h(t+u)\right\}\left\{\mathbb{K}_h(u-v)+\mathbb{K}_h(u+v)-\mathbb{K}_h(2M-u-v)\right\}\,du\\
&=\int_{u=0}^{M}\left\{K_h(t-u)-K_h(t+u)\right\}\left\{\mathbb{K}_h(u-v)+\mathbb{K}_h(u+v)-\mathbb{K}_h(2M-u-v)\right\}\,du,
\end{aligned}
$$

where we use $\mathbb{K}_h(-v)+\mathbb{K}_h(v)=1$ in the last equality (which follows from the symmetry of K). Furthermore,

$$\begin{aligned}
&\int_{u=0}^{M}\{K_h(t-u)-K_h(t+u)\}\{\mathbb{K}_h(u-v)+\mathbb{K}_h(u+v)-\mathbb{K}_h(2M-u-v)\}\,du\\
&=\int_{w=-1}^{t/h}K(w)\{\mathbb{K}_h(t-v-hw)+\mathbb{K}_h(t+v-hw)-1\}\,dw\\
&\qquad-\int_{w=t/h}^{1}K(w)\{\mathbb{K}_h(-t-v+hw)+\mathbb{K}_h(-t+v+hw)-1\}\,dw\\
&=\int_{w=-1}^{t/h}K(w)\{\mathbb{K}_h(t-v-hw)+\mathbb{K}_h(t+v-hw)-1\}\,dw\\
&\qquad+\int_{w=t/h}^{1}K(w)\{\mathbb{K}_h(t+v-hw)+\mathbb{K}_h(t-v-hw)-1\}\,dw\\
&=\int_{w=-1}^{1}K(w)\{\mathbb{K}_h(t-v-hw)+\mathbb{K}_h(t+v-hw)-1\}\,dw\\
&=\int K(w)\{\mathbb{K}_h(t-v-hw)+\mathbb{K}_h(t+v-hw)-K_h(2M-t-v-hw)\}\,dw\\
&=\widetilde{\mathbb{K}}_h(t-v)+\widetilde{\mathbb{K}}_h(t+v)-\widetilde{\mathbb{K}}_h(2M-t-v),
\end{aligned}$$

again using the relation $\mathbb{K}_h(x)+\mathbb{K}_h(-x)=1$. The case $t\in[M-h,M]$ is treated similarly.

B.2 The nonparametric bootstrap

To complete the notation introduced in Chapter 3, we suppose that the vectors $((Z_1,\ldots,Z_n),M_n), n=1,2,\ldots$ are defined on the product space $([0,M]\times\{0,1\})^\infty\times\mathbb{Z}_+^\infty,\mathcal{B},P_{ZM})$, where $\mathbb{Z}_+$ is the set of nonnegative integers and $\mathcal{B}$ is the collection of Borel sets, generated by the finite dimensional projections. We say that a real-valued function Γ_n defined on the joint probability space is of order $o_{\hat P}(1)$ in probability if for all $\epsilon,\eta>0$:

$$P^*\left(\hat P_n\{|\Gamma_n|>\epsilon\}>\eta\right)\to 0 \text{ as } n\to\infty,$$

where P^* denotes outer probability and $\hat P_n$ is the conditional probability measure w.r.t. the weights, given the sample $Z_1,\ldots,Z_n$.

B.2.1 Proof of Lemma 3.3.1

Before proving Lemma 3.3.1 we provide two technical lemmas.

Lemma B.2.1. *Let $\alpha > 0$. There exist constants $K_1, K_2 > 0$ such that, for each $j \geq 1$, $j \in \mathbb{N}$,*

$$\hat{P}_n\Bigg\{\exists y \in \big[(j-1)n^{-1/3}, jn^{-1/3}\big) : \\ \left|\int_{u\in(U_0(a),U_0(a)+y]} \{\delta - F_0(u)\}\, d\big(\hat{\mathbb{P}}_n - \mathbb{P}_n\big)(u,\delta)\right| \geq \alpha(j-1)^2 n^{-2/3}\Bigg\} \\ \leq K_1 \exp\left\{-K_2(j-1)^{3/2}\right\}, \tag{B.2.1}$$

in probability.

Likewise, there exist constants $K_1, K_2 > 0$ such that, for each $j \geq 1$, $j \in \mathbb{N}$,

$$\hat{P}_n\Bigg\{\exists y \in \big[-jn^{-1/3}, -(j-1)n^{-1/3}\big) : \\ \left|\int_{u\in(U_0(a)+y,U_0(a)]} \{\delta - F_0(u)\}\, d\big(\hat{\mathbb{P}}_n - \mathbb{P}_n\big)(u,\delta)\right| \geq \alpha(j-1)^2 n^{-2/3}\Bigg\} \\ \leq K_1 \exp\left\{-K_2(j-1)^{3/2}\right\}, \tag{B.2.2}$$

in probability.

Proof. We only prove (B.2.1), since the proof of (B.2.2) is similar. Let $\mathcal{F}_t$ be the (Vapnik-Cervonenkis) class of functions

$$\mathcal{F}_t = \left\{(\delta - F_0(v))1_{(U_0(a),U_0(a)+u]}(v) : u \in [0,t],\, \delta \in \{0,1\}\right\},$$

with envelope

$$F_t(v,\delta) = 1_{(U_0(a),U_0(a)+t]}(v), \qquad v \in [0,t].$$

To prove (B.2.1), we use that an exponential tail bound can be derived from a bounded Orlicz norm $\|\cdot\|_{P,\psi}$, i.e., when taking $\psi_1(x) = \exp(x) - 1$, for $x \geq 0$, we get, for $x > 0$ the inequality

$$P(|X| > x) \leq 2\exp\left\{-x/\|X\|_{P,\psi_1}\right\}, \tag{B.2.3}$$

where

$$\|X\|_{P,\psi_1} = \inf\left\{C > 0 : E\psi_1\left(\frac{|X|}{C}\right) \leq 1\right\}.$$

Using the second statement of Theorem 2.14.5 in van der Vaart and Wellner (1996), with $p = 1$, we get, the following inequality:

$$\left\|\left\|\sqrt{n}\left(\hat{\mathbb{P}}_n - \mathbb{P}_n\right)\right\|_{\mathcal{F}_t}^*\right\|_{\mathbb{P}_n,\psi_1} \lesssim \left\|\left\|\sqrt{n}\left(\hat{\mathbb{P}}_n - \mathbb{P}_n\right)\right\|_{\mathcal{F}_t}^*\right\|_{\mathbb{P}_n,1} + n^{-1/2}\{1+\log n\}\|F_t\|_{\mathbb{P}_n,\psi_1}, \tag{B.2.4}$$

where $\|\cdot\|_{\mathcal{F}_t}^*$ denotes the so-called measurable majorant of $\|\cdot\|_{\mathcal{F}_t}$ (see van der Vaart and Wellner (1996)). (Note that we use temporarily the "*" notation which is used for bootstrap variables in the rest of the paper.)
Furthermore, we have by the rightmost inequality of Theorem 2.14.1 of van der Vaart and Wellner (1996) that

$$\left\|\left\|\sqrt{n}\left(\hat{\mathbb{P}}_n-\mathbb{P}_n\right)\right\|_{\mathcal{F}_t}^*\right\|_{\mathbb{P}_n,1} \lesssim J\left(1,\mathcal{F}_t\right)\|F_t\|_{\mathbb{P}_n,2} ,$$

where $J(\delta,\mathcal{F}_t)$ is defined by

$$J(\delta,\mathcal{F}_t)=\sup_Q\int_0^\delta \sqrt{1+\log N\left(\varepsilon\|F\|_{Q,2},\mathcal{F}_t,L_2(Q)\right)}\,d\varepsilon,$$

and where the supremum is over all discrete probability measure Q with $\|F_t\|_{Q,2}>0$. Since $\mathcal{F}_t\subset\mathcal{F}_{M-U_0(a)}$ for all $t\in[0,M-U_0(a)]$, and since $\mathcal{F}_{M-U_0(a)}$ is a Vapnik-Cervonenkis class, $J(\delta,\mathcal{F}_t)$ is bounded by a fixed constant for all $t\in[0,M-U_0(a)]$, and we get that

$$\left\|\left\|\sqrt{n}\left(\hat{\mathbb{P}}_n-\mathbb{P}_n\right)\right\|_{\mathcal{F}_t}^*\right\|_{\mathbb{P}_n,1} \lesssim \|F_t\|_{\mathbb{P}_n,2} ,$$

uniformly for all $t\in[0,M-U_0(a)]$. Note that

$$\|F_t\|_{\mathbb{P}_n,2}^2=\int_{u\in U_0(a),U_0(a)+t]} d\mathbb{P}_n(u,\delta)=\int_{u\in U_0(a),U_0(a)+t]} d\mathbb{G}_n(u), \tag{B.2.5}$$

$t\in[U_0(a),M-U_0(a)]$. We next evaluate the second term on the right-hand side of (B.2.4). We have that

$$\int\psi_1\left(\frac{F_t(u,\delta)}{c}\right)d\mathbb{P}_n(u,\delta)=\left\{e^{1/c}-1\right\}\int 1_{(U_0(a),U_0(a)+t]}(u)\,d\mathbb{G}_n(u),$$

and

$$\begin{aligned}&\left\{e^{1/c}-1\right\}\int 1_{(U_0(a),U_0(a)+t]}(u)\,d\mathbb{G}_n(u)\le 1\\ &\iff c\ge\frac{1}{\log\left\{1+1/\int_{u\in U_0(a),U_0(a)+t]}d\mathbb{G}_n(u)\right\}}.\end{aligned}$$

Thus (B.2.4) becomes, using (B.2.5),

$$\begin{aligned}&\left\|\left\|\sqrt{n}\left(\hat{\mathbb{P}}_n-\mathbb{P}_n\right)\right\|_{\mathcal{F}_t}^*\right\|_{\mathbb{P}_n,\psi_1}\\ &\le c_1\left\{\int_{u\in U_0(a),U_0(a)+t]}d\mathbb{G}_n(u)\right\}^{1/2}+\frac{1+\log n}{n^{1/2}\log\left\{1+1/\int_{u\in U_0(a),U_0(a)+t]}d\mathbb{G}_n(u)\right\}},\end{aligned} \tag{B.2.6}$$

for a constant $c_1 > 0$. If $t \geq Kn^{-1/3}$ we get for the second term in probability,

$$\frac{1+\log n}{n^{1/2}\log\left\{1+1/\int_{u\in U_0(a),U_0(a)+t]} d\mathbb{G}_n(u)\right\}} \ll c_1\left\{\int_{u\in U_0(a),U_0(a)+t]} d\mathbb{G}_n(u)\right\}^{1/2}.$$

We have:

$$\begin{aligned}
&\int_{u\in[U_0(a),U_0(a)+t]} d\mathbb{G}_n(u)\\
&= \int_{u\in[U_0(a),U_0(a)+t]} dG(u) + \int_{u\in[U_0(a),U_0(a)+t]} d\big(\mathbb{G}_n - G\big)(u)\\
&= \int_{u\in[U_0(a),U_0(a)+t]} dG(u) + O_p\left(n^{-1/2}\right) = O(t) + O_p\left(n^{-1/2}\right)\\
&= O(t) + O_{\hat P}\left(n^{-1/2}\right),
\end{aligned}$$

in probability (since a term defined only on the probability space $(\mathcal{X},\mathcal{A},P)$ of order $O_p(1)$ is also of order $O_{\hat P}(1)$ in probability). So we obtain, for $j \geq K$ in probability, conditioning on $(T_1,\Delta_1),(T_2,\Delta_2),\ldots$ using the inequality on Orlicz norms on p. 96 or 239 of van der Vaart and Wellner (1996):

$$\begin{aligned}
&\hat P_n\Bigg\{\exists y \in \big[(j-1)n^{-1/3}, jn^{-1/3}\big) :\\
&\qquad\left|\int_{u\in(U_0(a),U_0(a)+y]}\{\delta - F_0(u)\}\, d\big(\hat{\mathbb{P}}_n - \mathbb{P}_n\big)(u,\delta)\right| \geq \alpha(j-1)^2 n^{-2/3}\Bigg\}\\
&= \hat P_n\Bigg\{\exists y \in \big[(j-1)n^{-1/3}, jn^{-1/3}\big) :\\
&\qquad\sqrt{n}\left|\int_{u\in(U_0(a),U_0(a)+y]}\{\delta - F_0(u)\}\, d\big(\hat{\mathbb{P}}_n - \mathbb{P}_n\big)(u,\delta)\right| \geq \alpha(j-1)^2 n^{-1/6}\Bigg\}\\
&\leq 2\exp\left\{-m(j-1)^2 n^{-1/6}\Big/\left\|\left\|\sqrt{n}\left(\hat{\mathbb{P}}_n - \mathbb{P}_n\right)\right\|^*_{\mathcal{F}_{jn^{-1/3}}}\right\|_{\mathbb{P}_n,\psi_1}\right\}\\
&\leq 2\exp\left\{-c_2 m(j-1)^{3/2}\right\},
\end{aligned}$$

for some $c_2 > 0$. This proves the statement. □

Lemma B.2.2. *For each $\varepsilon > 0$ and $x \in [0, M - U_0(a)]$,*

$$\left|\int_{u\in(U_0(a),U_0(a)+x]}\{\delta - F_0(u)\}\, d\big(\mathbb{P}_n - P\big)(u,\delta)\right| \leq \varepsilon x^2 + O_p\left(n^{-2/3}\right).$$

Proof. As in the proof of Lemma B.2.1, we consider the Vapnik-Cervonenkis collection of functions

$$\mathcal{F}_t = \left\{(\delta - F_0(v))1_{(U_0(a),U_0(a)+u]}(v) : u \in [0,t],\, \delta \in \{0,1\}\right\},$$

with envelope

$$F_t(v,\delta) = 1_{(U_0(a),U_0(a)+t]}(v), \qquad v \in [0,t].$$

We have, using Theorem 2.14.1 of van der Vaart and Wellner (1996), that

$$E_X\left\{\sup_{f\in\mathcal{F}_t} |\mathbb{P}_n - P|\,(f)\right\}^2 \le Kn^{-1}\,\|F_t\|_{P,2}^2\,, \tag{B.2.7}$$

for some $K > 0$. Since

$$\|F_t\|_{P,2}^2 = \int_{u\in U_0(a),U_0(a)+t]} dP(u,\delta) = \int_{u\in U_0(a),U_0(a)+t]} dG(u) = O(t),$$

for $t \in [U_0(a), M - U_0(a)]$, we get, by Markov's inequality,

$$P\left\{n^{2/3}\left|\int_{u\in(U_0(a),U_0(a)+jn^{-1/3}]}\{\delta - F_0(u)\}\,d\big(\mathbb{P}_n - P\big)(u,\delta)\right| > A + \varepsilon(j-1)^2\right\}$$
$$\le Kj/\left\{A + \varepsilon(j-1)^2\right\}^2.$$

The result now easily follows, see e.g. Kim and Pollard (1990). p. 201. □

As a consequence of Lemma B.2.1 and Lemma B.2.2 we get the following result.

Lemma B.2.3. *Let $\hat{V}_n$ and $\hat{\hat{V}}_n$ be defined by*

$$\hat{V}_n(t) = \int_{u\in[0,t]} \delta\, d\hat{\mathbb{P}}_n(u,\delta), \qquad \hat{\hat{V}}_n(t) = \int_{u\in[0,t]} F_0(u)\, d\hat{\mathbb{G}}_n(u), \qquad t \in [0,M]. \tag{B.2.8}$$

where the process $\hat{\mathbb{G}}_n$ is defined in (3.3.1), and let $\hat{D}_n = \hat{V}_n - \hat{\hat{V}}_n$. Then there exist constants $K_1, K_2 > 0$ such that, for each $j \ge 1$, $j \in \mathbb{N}$,

$$\hat{P}_n\Big\{\exists y \in \big[(j-1)n^{-1/3}, jn^{-1/3}\big) : \hat{D}_n(U_0(a)+y) - \hat{D}_n(U_0(a))$$
$$\le -\int_{U_0(a)}^{U_0(a)+y}\{F_0(u) - F_0(U_0(a))\}\, d\hat{\mathbb{G}}_n(u)\Big\}$$
$$\le K_1 \exp\left\{-K_2(j-1)^{3/2}\right\}, \tag{B.2.9}$$

in probability. Likewise, there exist constants $K_1, K_2 > 0$ such that, for each $j \geq 1,\ j \in \mathbb{N}$,

$$\hat{P}_n\left\{\exists y \in \bigl[-jn^{-1/3}, -(j-1)n^{-1/3}\bigr) : \hat{D}_n(U_0(a)+y) - \hat{D}_n(U_0(a)) \right.$$
$$\left. \leq -\int_{U_0(a)+y}^{U_0(a)} \bigl\{F_0(u) - F_0(U_0(a))\bigr\}\, d\hat{\mathbb{G}}_n(u)\right\}$$
$$\leq K_1 \exp\left\{-K_2(j-1)^{3/2}\right\}, \tag{B.2.10}$$

in probability.

Proof. We again only prove (B.2.1), since the proof of (B.2.2) is similar. First note that

$$\hat{P}_n\left\{\exists y \in \bigl[(j-1)n^{-1/3}, jn^{-1/3}\bigr) : \hat{D}_n(U_0(a)+y) - \hat{D}_n(U_0(a)) \right.$$
$$\left. \leq -\int_{U_0(a)}^{U_0(a)+y} \bigl\{F_0(u) - F_0(U_0(a))\bigr\}\, d\hat{\mathbb{G}}_n(u)\right\}$$
$$\leq \hat{P}_n\left\{\exists y \in \bigl[(j-1)n^{-1/3}, jn^{-1/3}\bigr) : \left|\hat{D}_n(U_0(a)+y) - \hat{D}_n(U_0(a))\right| \right.$$
$$\left. \geq \int_{U_0(a)}^{U_0(a)+y} \bigl\{F_0(u) - F_0(U_0(a))\bigr\}\, d\hat{\mathbb{G}}_n(u)\right\}.$$

Furthermore,

$$\int_{U_0(a)}^{U_0(a)+y} \bigl\{F_0(u) - F_0(U_0(a))\bigr\}\, d\hat{\mathbb{G}}_n(u)$$
$$= \int_{U_0(a)}^{U_0(a)+y} \bigl\{F_0(u) - F_0(U_0(a))\bigr\}\, d\mathbb{G}_n(u)$$
$$+ \int_{U_0(a)}^{U_0(a)+y} \bigl\{F_0(u) - F_0(U_0(a))\bigr\}\, d\bigl(\hat{\mathbb{G}}_n - \mathbb{G}_n\bigr)(u)$$
$$= \int_{U_0(a)}^{U_0(a)+y} \bigl\{F_0(u) - F_0(U_0(a))\bigr\}\, dG(u)$$
$$+ \int_{U_0(a)}^{U_0(a)+y} \bigl\{F_0(u) - F_0(U_0(a))\bigr\}\, d\bigl(\mathbb{G}_n - G\bigr)(u)$$
$$+ \int_{U_0(a)}^{U_0(a)+y} \bigl\{F_0(u) - F_0(U_0(a))\bigr\}\, d\bigl(\hat{\mathbb{G}}_n - \mathbb{G}_n\bigr)(u), \tag{B.2.11}$$

and for the dominant term on the right-hand side we get that

$$\int_{U_0(a)}^{U_0(a)+y} \bigl\{F_0(u) - F_0(U_0(a))\bigr\}\, dG(u) \geq m_0 \int_{U_0(a)}^{U_0(a)+y} \{u - U_0(a)\}\, dG(u)$$
$$\geq m_0 m_1 \int_{U_0(a)}^{U_0(a)+y} \{u - U_0(a)\}\, du = \tfrac{1}{2} m_0 m_1 \{y - U_0(a)\}^2,$$

where $m_0 = \inf_{u\in[U_0(a),M]} f_0(u)$ and $m_1 = \inf_{u\in[U_0(a),M]} g(u)$. We therefore consider the probability

$$\hat{P}_n\Bigl\{\exists y \in \bigl[(j-1)n^{-1/3}, jn^{-1/3}\bigr) : \Bigl|\hat{D}_n(U_0(a)+y) - \hat{D}_n(U_0(a))\Bigr| \geq m(j-1)^2 n^{-2/3}\Bigr\}, \quad \text{(B.2.12)}$$

where

$$m = \tfrac{1}{2} \min\left\{ \inf_{u\in[t_0,M]} f_0(u), \inf_{u\in[t_0,M]} g(u)\right\}.$$

We also have that

$$\hat{D}_n(U_0(a)+y) - \hat{D}_n(U_0(a)) = \int_{u\in(U_0(a),U_0(a)+y]} \{\delta - F_0(u)\}\, d\hat{\mathbb{P}}_n(u,\delta)$$
$$= \int_{u\in(U_0(a),U_0(a)+y]} \{\delta - F_0(u)\}\, d\bigl(\hat{\mathbb{P}}_n - P\bigr)(u,\delta)$$
$$= \int_{u\in(U_0(a),U_0(a)+y]} \{\delta - F_0(u)\}\, d\bigl(\hat{\mathbb{P}}_n - \mathbb{P}_n\bigr)(u,\delta)$$
$$+ \int_{u\in(U_0(a),U_0(a)+y]} \{\delta - F_0(u)\}\, d\bigl(\mathbb{P}_n - P\bigr)(u,\delta).$$

By Lemma B.2.2, we may assume that for $x \in [0, M - U_0(a)]$,

$$\left|\int_{u\in(U_0(a),U_0(a)+x]} \{\delta - F_0(u)\}\, d\bigl(\mathbb{P}_n - P\bigr)(u,\delta)\right| \leq \varepsilon x^2 + Kn^{-2/3}, \qquad \text{(B.2.13)}$$

for some $K > 0$ and $0 < \varepsilon < m/2$. Considering sequences $X = (T_1,\Delta_1),(T_2,\Delta_2)\ldots$, satisfying (B.2.13), we get that

$$\hat{P}_n\Bigl\{\exists y \in \bigl[(j-1)n^{-1/3}, jn^{-1/3}\bigr) : \Bigl|\hat{D}_n(U_0(a)+y) - \hat{D}_n(U_0(a))\Bigr| \geq m(j-1)^2 n^{-2/3}\Bigr\}$$
$$\leq \hat{P}_n\left\{\exists y \in \bigl[(j-1)n^{-1/3}, jn^{-1/3}\bigr) : \left|\int_{u\in(U_0(a),U_0(a)+y]} \{\delta - F_0(u)\}\, d\bigl(\hat{\mathbb{P}}_n - \mathbb{P}_n\bigr)(u,\delta)\right| \geq \tfrac{1}{2} m(j-1)^2 n^{-2/3}\right\}$$
$$\leq K_1 \exp\left\{-K_2(j-1)^{3/2}\right\},$$

with probability tending to one, using Lemma B.2.1. □

We now prove Lemma 3.1.1.

Proof of Lemma 3.1.1. Suppose that $n^{1/3}|\hat{U}_n(a) - U_0(a)| > x$ for some $x > 0$, then there exists a y such that, $n^{1/3}\,|y - U_0(a)| > x$ and $\hat{V}_n(y) - a\hat{\mathbb{G}}_n(y) \le \hat{V}_n(U_0(a)) - a\hat{\mathbb{G}}_n(U_0(a))$. Hence,

$$\begin{aligned}
&\hat{P}_n\left\{n^{1/3}\left|\hat{U}_n(a) - U_0(a)\right| \ge x\right\}\\
&\le \hat{P}_n\Bigg(\inf_{y - U_0(a) \ge n^{-1/3}x} \hat{D}_n(y) - \hat{D}_n(U_0(a))\\
&\qquad\qquad\qquad \le -\int_{U_0(a)}^{y}\left\{F_0(u) - F_0(U_0(a))\right\} d\hat{\mathbb{G}}_n(u)\Bigg)\\
&\le \sum_{j=i}^{\infty}\hat{P}_n\Bigg(\exists y \in \left[(j-1)n^{-1/3}, jn^{-1/3}\right) : \hat{D}_n(U_0(a)+y) - \hat{D}_n(U_0(a))\\
&\qquad\qquad\qquad \le -\int_{U_0(a)}^{U_0(a)+y}\left\{F_0(u) - F_0(U_0(a))\right\} d\hat{\mathbb{G}}_n(u)\Bigg),
\end{aligned}$$

where $x \in [(i-1)n^{-1/3}, in^{-1/3}]$. By Lemma B.2.3, this is bounded above by

$$\begin{aligned}
&\sum_{j=i}^{\infty} K_1 \exp\left\{K_2(j-1)^{3/2}\right\}\\
&= K_1 \exp\left\{-K_2(i-1)^{3/2}\right\}\sum_{j=i}^{\infty}\exp\left\{-K_2[(j-1)^{3/2} - (i-1)^{3/2}]\right\}\\
&\le K_1' \exp\left\{K_2'(i-1)^{3/2}\right\},
\end{aligned}$$

for constants $K_1, K_1', K_2, K_2' > 0$. □

B.2.2 Asymptotic normality of the bootstrapped SMLE

In this section, we prove (3.3.2). Define the function $\psi_{t,h}$ as in (B.1.12) and denote the points of jump of the nonparametric bootstrapped MLE $\hat{F}_n^*$ by $\hat{\tau}_1, \ldots, \hat{\tau}_m$ and define the piecewise constant function $\bar{\psi}_{t,h}$ with only jumps at $\hat{\tau}_1, \ldots, \hat{\tau}_m$ by

$$\bar{\psi}_{t,h}(u) = \begin{cases} \psi_{t,h}(\hat{\tau}_i), & \text{if } F_0(u) > \hat{F}_n^*(\hat{\tau}_i),\ u \in [\hat{\tau}_i, \hat{\tau}_{i+1}),\\ \psi_{t,h}(s), & \text{if } F_0(u) = \hat{F}_n^*(s),\ \text{for some } s \in [\hat{\tau}_i, \hat{\tau}_{i+1}),\\ \psi_{t,h}(\hat{\tau}_{i+1}), & \text{if } \tilde{F}_0(u) < \hat{F}_n^*(\tau_i),\ u \in [\hat{\tau}_i, \hat{\tau}_{i+1}). \end{cases}$$

By the convex minorant interpretation of $\hat{F}_n^*$, we have that

$$\int \bar{\psi}_{t,h}(u)(\delta - \hat{F}_n^*(u))d\hat{\mathbb{P}}_n(u,\delta) = 0,$$

(see the discussion of the SMLE in Groeneboom and Jongbloed (2014), p. 332).
We can write:

$$\begin{aligned}
\tilde{F}_{nh}^*(t) &= \int \mathbb{K}_h(t-u)\, d\hat{F}_n^*(u) \\
&= \int \mathbb{K}_h(t-u)\, d(\hat{F}_n^* - F_0)(u) + \int \mathbb{K}_h(t-u)\, dF_0(u) \\
&= \int \psi_{t,h}(u)\, \left\{\hat{F}_n^*(u) - F_0(u)\right\} dG(u) + \int \mathbb{K}_h(t-u)\, dF_0(u) \\
&= \int \psi_{t,h}(u) \left\{F_0(u) - \hat{F}_n^*(u)\right\} d(\hat{\mathbb{G}}_n - G)(u) + \int \psi_{t,h}(u)\, \{\delta - F_0(u)\}\, d\hat{\mathbb{P}}_n(u,\delta) \\
&\quad + \int \{\psi_{t,h}(u) - \bar{\psi}_{t,h}(u)\} \left\{\hat{F}_n^*(u) - \delta\right\} d\hat{\mathbb{P}}_n(u,\delta) + \int \mathbb{K}_h(t-u)\, dF_0(u) \\
&= \tilde{F}_{nh}^{(toy)*}(t) + \int \psi_{t,h}(u) \left\{F_0(u) - \hat{F}_n^*(u)\right\} d(\hat{\mathbb{G}}_n - G)(u,\delta) \\
&\quad + \int \{\psi_{t,h}(u) - \bar{\psi}_{t,h}(u)\} \left\{\hat{F}_n^*(u) - \delta\right\} d\hat{\mathbb{P}}_n(u,\delta) \\
&= \tilde{F}_{nh}^{(toy)*}(t) + A_I + A_{II}.
\end{aligned}$$

We first evaluate A_I and show that this term is $o_{\hat{P}}(n^{-2/5})$ in probability, we have that

$$\begin{aligned}
A_I &= \int \psi_{t,h}(u) \left\{F_0(u) - \hat{F}_n^*(u)\right\} d(\hat{\mathbb{G}}_n - G)(u,\delta) \\
&= \int \psi_{t,h}(u) \left\{F_0(u) - \hat{F}_n^*(u)\right\} d(\hat{\mathbb{G}}_n - \mathbb{G}_n)(u,\delta) \\
&\quad + \int \psi_{t,h}(u) \left\{F_0(u) - \hat{F}_n^*(u)\right\} d(\mathbb{G}_n - G)(u,\delta)
\end{aligned}$$

An argument similar to that of Lemma A.7 in Groeneboom et al. (2010) shows that

$$\int \psi_{t,h}(u) \left\{F_0(u) - \hat{F}_n^*(u)\right\} d(\mathbb{G}_n - G)(u,\delta) = o_p(n^{-2/5}),$$

and hence,

$$\int \psi_{t,h}(u) \left\{F_0(u) - \hat{F}_n^*(u)\right\} d(\mathbb{G}_n - G)(u,\delta) = o_{\hat{P}}(n^{-2/5}),$$

in probability. Similarly to the proof of Lemma A.7 in Groeneboom et al. (2010), we can also show that

$$\int \psi_{t,h}(u) \left\{F_0(u) - \hat{F}_n^*(u)\right\} d(\hat{\mathbb{G}}_n - \mathbb{G}_n)(u,\delta) = o_{\hat{P}}(n^{-2/5}), \qquad \text{(B.2.14)}$$

in probability, such that

$$A_I = o_{\hat{P}}(n^{-2/5}) \qquad \text{in probability.}$$

We now study the term A_{II}. Using the same inequality for $\psi_{t,h} - \bar{\psi}_{t,h}$ as used in the second display after (11.49) on p. 333 of Groeneboom and Jongbloed (2014), we get for some constant $C > 0$ that

$$\left|\bar{\psi}_{t,h}(u) - \psi_{t,h}(u)\right| \le Ch^{-2}\left|\hat{F}_n^*(u) - F_0(u)\right|, \tag{B.2.15}$$

for all u such that f_0 is positive and continuous in a neighborhood around u. We decompose the term A_{II} as follows:

$$\begin{aligned} A_{II} = &\int \left\{\bar{\psi}_{t,h}(u) - \psi_{t,h}(u)\right\}\left\{\hat{F}_n^*(u) - F_0(u)\right\} d\hat{\mathbb{P}}_n(u,\delta) \\ &+ \int \left\{\bar{\psi}_{t,h}(u) - \psi_{t,h}(u)\right\}\left\{F_0(u) - \delta\right\} d\hat{\mathbb{P}}_n(u,\delta). \end{aligned} \tag{B.2.16}$$

For the first term on the right-hand side of the above display we write:

$$\begin{aligned} &\int \left\{\bar{\psi}_{t,h}(u) - \psi_{t,h}(u)\right\}\left\{\hat{F}_n^*(u) - F_0(u)\right\} d\hat{\mathbb{P}}_n(u,\delta) \\ &\quad = \int \left\{\bar{\psi}_{t,h}(u) - \psi_{t,h}(u)\right\}\left\{\hat{F}_n^*(u) - F_0(u)\right\} d(\hat{\mathbb{P}}_n - \mathbb{P}_n)(u,\delta) \\ &\qquad + \int \left\{\bar{\psi}_{t,h}(u) - \psi_{t,h}(u)\right\}\left\{\hat{F}_n^*(u) - F_0(u)\right\} d\mathbb{P}_n(u,\delta) \\ &\quad \le \int \left\{\bar{\psi}_{t,h}(u) - \psi_{t,h}(u)\right\}\left\{\hat{F}_n^*(u) - F_0(u)\right\} d(\hat{\mathbb{P}}_n - \mathbb{P}_n)(u,\delta) \\ &\qquad + Ch^{-2}\int_{t-h}^{t+h}\left\{\hat{F}_n^*(u) - F_0(u)\right\}^2 d\mathbb{P}_n(u,\delta), \end{aligned} \tag{B.2.17}$$

where we use (B.2.15) in the last inequality. The first term in the display above is $o_{\hat{P}}(n^{-2/5})$ in probability by (B.2.14) and (B.2.15). Since

$$\hat{E}_n\left\{\hat{F}_n^*(t) - F_0(t)\right\}^2 < Kn^{-2/3} \qquad \forall t \in (0, M),$$

in probability, we have by Markov's inequality and Fubini's theorem that

$$\int_{t-h}^{t+h}\left\{\hat{F}_n^*(u) - F_0(u)\right\}^2 d\mathbb{P}_n(u,\delta) = O_{\hat{P}}\left(hn^{-2/3}\right) \text{ in probability.} \tag{B.2.18}$$

Hence, for $h \asymp n^{-1/5}$, we get for the second term in (B.2.17) that

$$\begin{aligned} &Ch^{-2}\int_{t-h}^{t+h}\left\{\hat{F}_n^*(u) - F_0(u)\right\}^2 d\mathbb{P}_n(u,\delta) \\ &\quad = O_{\hat{P}}\left(h^{-1}n^{-2/3}\right) = O_{\hat{P}}\left(n^{-7/15}\right) = o_{\hat{P}}\left(n^{-2/5}\right), \end{aligned}$$

in probability. For the second term of (B.2.16) we have that

$$\begin{aligned}\int \left\{\bar\psi_{t,h}(u)-\psi_{t,h}(u)\right\}\left\{F_0(u)-\delta\right\} d\hat{\mathbb{P}}_n(u,\delta) \\ = \int \left\{\bar\psi_{t,h}(u)-\psi_{t,h}(u)\right\}\left\{F_0(u)-\delta\right\} d(\hat{\mathbb{P}}_n-\mathbb{P}_n)(u,\delta) \\ + \int \left\{\bar\psi_{t,h}(u)-\psi_{t,h}(u)\right\}\left\{F_0(u)-\delta\right\} d\big(\mathbb{P}_n-P\big)(u,\delta).\end{aligned}$$

Similar to the arguments used in the treatment of term A_I above, we get by using again arguments similar to that of Lemma A.7 in Groeneboom et al. (2010) that

$$\int \left\{\bar\psi_{t,h}(u)-\psi_{t,h}(u)\right\}\left\{F_0(u)-\delta\right\} d\big(\hat{\mathbb{P}}_n-\mathbb{P}_n\big)(u,\delta) = o_{\hat P}(n^{-2/5}),$$

and

$$\int \left\{\bar\psi_{t,h}(u)-\psi_{t,h}(u)\right\}\left\{F_0(u)-\delta\right\} d\big(\mathbb{P}_n-P\big)(u,\delta) = o_{\hat P}(n^{-2/5}),$$

in probability.

B.2.3 Asymptotic normality of the SSE

In this section we give a road map for the proof of the bootstrap validity in the current status linear regression model given in (3.6.3). We assume that the assumptions stated in Theorem 2.4.1 hold. Since the proof is very similar to the proof of Theorem 2.4.1, we leave the details to the interested reader. Consider the bootstrap score function

$$\hat\psi_{n\epsilon}(\beta) = \int_{\hat F^*_{n,\boldsymbol\alpha}(t-\boldsymbol\alpha^T\boldsymbol x)\in[\epsilon,1-\epsilon]} x\{\delta-\hat F^*_{n,\boldsymbol\alpha}(t-\boldsymbol\alpha^T\boldsymbol x)\}\, d\hat{\mathbb{P}}_n(\boldsymbol x,t,\delta), \tag{B.2.19}$$

for some fixed truncation parameter $\epsilon\in(0,1/2)$.
The main idea is to show that

$$\begin{aligned}\hat\psi_{n\epsilon}(\hat{\boldsymbol\alpha}^*_n) = \boldsymbol A(\hat{\boldsymbol\alpha}^*_n-\boldsymbol\alpha_0) + \int_{F_0(t-\boldsymbol\alpha_0^T\boldsymbol x)\in[\epsilon,1-\epsilon]} \{\boldsymbol x-\mathbb{E}(\boldsymbol X|T-\boldsymbol\alpha_0^T\boldsymbol X=t-\boldsymbol\alpha_0^T\boldsymbol x)\} \\ \cdot\{\delta-F_0(t-\boldsymbol\alpha_0' x)\}\, d(\hat{\mathbb{P}}_n-\mathbb{P}_n)(\boldsymbol x,t,\delta) \\ + \int_{F_0(t-\boldsymbol\alpha_0^T\boldsymbol x)\in[\epsilon,1-\epsilon]} \{\boldsymbol x-\mathbb{E}(\boldsymbol X|T-\boldsymbol\alpha_0^T\boldsymbol X=t-\boldsymbol\alpha_0^T\boldsymbol x)\} \\ \cdot\{\delta-F_0(t-\boldsymbol\alpha_0' x)\}\, d(\mathbb{P}_n-P)(\boldsymbol x,t,\delta) \\ + o_{\hat P}(n^{-1/2}+(\hat{\boldsymbol\alpha}^*_n-\boldsymbol\alpha_0)),\end{aligned} \tag{B.2.20}$$

in probability. As in the proof of Theorem 2.4.1 given in Appendix A, we can work with the definition

$$\hat\psi_{n\epsilon}(\hat{\boldsymbol\alpha}^*_n) = \boldsymbol 0,$$

for the score estimator $\hat{\boldsymbol{\alpha}}_n^*$. Since by the proof of Theorem 2.4.1,

$$\begin{aligned}
&-\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) \\
&\quad = \sqrt{n}\int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X} = t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\} \\
&\qquad\qquad\qquad\qquad\qquad \cdot \{\delta - F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\}\, d(\mathbb{P}_n - P)(\boldsymbol{x},t,\delta) \\
&\qquad + o_p(1+\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)),
\end{aligned}$$

we get that,

$$\begin{aligned}
&-\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\alpha}}_n^* - \hat{\boldsymbol{\alpha}}_n) \\
&\quad = \sqrt{n}\int_{F_0(t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|T-\boldsymbol{\alpha}_0^T\boldsymbol{X} = t-\boldsymbol{\alpha}_0^T\boldsymbol{x})\} \\
&\qquad\qquad\qquad\qquad\qquad \cdot \{\delta - F_0(t-\boldsymbol{\alpha}_0' x)\}\, d(\hat{\mathbb{P}}_n - \mathbb{P}_n)(\boldsymbol{x},t,\delta) \\
&\qquad + o_{\hat{P}}(1+\sqrt{n}(\hat{\boldsymbol{\alpha}}_n^* - \boldsymbol{\alpha}_0)).
\end{aligned}$$

The validity of the bootstrap then follows by the arguments given in Section 3.6. Very important in the proof of (B.2.20) is the conditional bootstrapped L_2-result,

$$\sup_{\boldsymbol{\alpha}} \int \left\{\hat{F}_{n,\boldsymbol{\alpha}}^*(t-\boldsymbol{\alpha}^T\boldsymbol{x}) - F_{\boldsymbol{\alpha}}(t-\boldsymbol{\alpha}^T\boldsymbol{x})\right\}^2 d\mathbb{P}_n(\boldsymbol{x},t,\delta) = O_{\hat{P}}\left(n^{-2/3}\right), \qquad \text{(B.2.21)}$$

in probability, where $F_{\boldsymbol{\alpha}}$ is defined in (2.3.1).

Let $\bar{\phi}_{\hat{\boldsymbol{\alpha}}_n^*,\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}}$ be a (random) piecewise constant version of $\phi_{\hat{\boldsymbol{\alpha}}_n^*}$, where

$$\phi_{\boldsymbol{\alpha}} = \mathbb{E}\left\{\boldsymbol{X}|T-\boldsymbol{\alpha}^T\boldsymbol{X} = u\right\},$$

and where, for a piecewise constant distribution function F with finitely many jumps at $\tau_1 < \tau_2 < \ldots$, the function $\bar{\phi}_{\boldsymbol{\alpha},F}$ is defined in the following way.

$$\bar{\phi}_{\boldsymbol{\alpha},F}(u) = \begin{cases} \phi_{\boldsymbol{\alpha}}(\tau_i), & \text{if } F_{\boldsymbol{\alpha}}(u) > F(\tau_i),\ u\in[\tau_i,\tau_{i+1}), \\ \phi_{\boldsymbol{\alpha}}(s), & \text{if } F_{\boldsymbol{\alpha}}(u) = F(s),\ \text{for some } s\in[\tau_i,\tau_{i+1}), \\ \phi_{\boldsymbol{\alpha}}(\tau_{i+1}), & \text{if } F_{\boldsymbol{\alpha}}(u) < F(\tau_i),\ u\in[\tau_i,\tau_{i+1}). \end{cases} \qquad \text{(B.2.22)}$$

Similar to the proof of Theorem 2.4.1, we get that

$$\|\phi_{\hat{\boldsymbol{\alpha}}_n^*}(u) - \bar{\phi}_{\hat{\boldsymbol{\alpha}}_n^*,\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}}(u)\| \le K|\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}(u) - F_{\hat{\boldsymbol{\alpha}}_n^*}(u)|,$$

for some constant $K > 0$ not depending on $\boldsymbol{\alpha}$. By the definition of the MLE $\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}$ as the slope of the greatest convex minorant of the corresponding cusum diagram, we can write:

$$
\begin{aligned}
\hat{\psi}_{n\epsilon}(\hat{\boldsymbol{\alpha}}_n^*) = &\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{x - \phi_{\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\right\} \\
&\qquad\qquad \cdot \left\{\delta - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\right\} d\hat{\mathbb{P}}_n(\boldsymbol{x},t,\delta) \\
&+ \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\phi_{\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x}) - \bar{\phi}_{\hat{\boldsymbol{\alpha}}_n^*,\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\right\} \\
&\qquad\qquad \cdot \left\{\delta - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\right\} d\hat{\mathbb{P}}_n(\boldsymbol{x},t,\delta) \\
= &\, I + II,
\end{aligned}
$$

For the second term, we have that

$$
\begin{aligned}
II = &\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\phi_{\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x}) - \bar{\phi}_{\hat{\boldsymbol{\alpha}}_n^*,\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\right\} \\
&\qquad\qquad \cdot \left\{\delta - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\right\} d(\hat{\mathbb{P}}_n - \mathbb{P}_n)(\boldsymbol{x},t,\delta) \\
&+ \int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{\phi_{\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x}) - \bar{\phi}_{\hat{\boldsymbol{\alpha}}_n^*,\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\right\} \\
&\qquad\qquad \cdot \left\{\delta - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\right\} d\mathbb{P}_n(\boldsymbol{x},t,\delta) \\
= &\, II_a + II_b
\end{aligned}
$$

It is shown in the proof of Theorem 2.4.1 that

$$II_b = o_p(n^{-1/2} + (\hat{\boldsymbol{\alpha}}_n^* - \boldsymbol{\alpha}_0)),$$

and therefore

$$II_b = o_{\hat{P}}(n^{-1/2} + (\hat{\boldsymbol{\alpha}}_n^* - \boldsymbol{\alpha}_0)) \text{ in probability.}$$

Using similar arguments as in in the proof of Theorem 2.4.1 we can also show that

$$II_a = o_{\hat{P}}(n^{-1/2}) \text{ in probability.}$$

Hence, we get that'12

$$
\begin{aligned}
\hat{\psi}_{n\epsilon}(\hat{\boldsymbol{\alpha}}_n^*) = &\int_{\hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\in[\epsilon,1-\epsilon]} \left\{x - \phi_{\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\right\} \\
&\qquad\qquad \cdot \left\{\delta - \hat{F}_{n,\hat{\boldsymbol{\alpha}}_n^*}(t-(\hat{\boldsymbol{\alpha}}_n^*)^T\boldsymbol{x})\right\} d\hat{\mathbb{P}}_n(\boldsymbol{x},t,\delta) \\
&+ o_{\hat{P}}(n^{-1/2} + (\hat{\boldsymbol{\alpha}}_n^* - \boldsymbol{\alpha}_0)),
\end{aligned}
$$

in probability. We now write

$$\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\in[\epsilon,1-\epsilon]} \big\{x-\phi_{\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\big\}\big\{\delta-\hat F_{n,\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\big\}\,d\hat{\mathbb P}_n(\boldsymbol x,t,\delta)$$
$$=\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\in[\epsilon,1-\epsilon]} \big\{x-\phi_{\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\big\}$$
$$\cdot\big\{\delta-\hat F_{n,\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\big\}\,d(\hat{\mathbb P}_n-\mathbb P_n)(\boldsymbol x,t,\delta)$$
$$+\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\in[\epsilon,1-\epsilon]} \big\{x-\phi_{\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\big\}$$
$$\cdot\big\{\delta-\hat F_{n,\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\big\}\,d\mathbb P_n(\boldsymbol x,t,\delta)$$

It follows from the proof of Theorem 2.4.1 that there exists a random variable R_n of order $o_p(n^{-1/2}+\hat{\boldsymbol\alpha}_n^*-\boldsymbol\alpha_0)$ (and hence of order $o_{\hat P}(n^{-1/2}+\hat{\boldsymbol\alpha}_n^*-\boldsymbol\alpha_0)$ in probability) such that

$$\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\in[\epsilon,1-\epsilon]} \big\{x-\phi_{\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\big\}\big\{\delta-\hat F_{n,\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\big\}\,d\mathbb P_n(\boldsymbol x,t,\delta)$$
$$=\int_{F_0(t-\boldsymbol\alpha_0^T\boldsymbol x)\in[\epsilon,1-\epsilon]}\Big\{x-\phi_0(t-\boldsymbol\alpha_0^T\boldsymbol x)\Big\}\Big\{\delta-F_0(t-\boldsymbol\alpha_0^T\boldsymbol x)\Big\}\,d(\mathbb P_n-P)(\boldsymbol x,t,\delta)$$
$$+\psi_{1,\epsilon}'(\boldsymbol\alpha_0)(\hat{\boldsymbol\alpha}_n^*-\boldsymbol\alpha_0)+R_n. \tag{B.2.23}$$

where $\phi_0\equiv\phi_{\boldsymbol\alpha_0}$. Therefore, (B.2.20) follows if we can show that

$$\int_{\hat F_{n,\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\in[\epsilon,1-\epsilon]} \big\{x-\phi_{\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\big\}$$
$$\cdot\big\{\delta-\hat F_{n,\hat{\boldsymbol\alpha}_n^*}(t-(\hat{\boldsymbol\alpha}_n^*)^T\boldsymbol x)\big\}\,d(\hat{\mathbb P}_n-\mathbb P_n)(\boldsymbol x,t,\delta)$$
$$=\int_{F_0(t-\boldsymbol\alpha_0^T\boldsymbol x)\in[\epsilon,1-\epsilon]}\Big\{x-\phi_0(t-\boldsymbol\alpha_0^T\boldsymbol x)\Big\}\Big\{\delta-F_0(t-\boldsymbol\alpha_0^T\boldsymbol x)\Big\}\,d(\hat{\mathbb P}_n-\mathbb P_n)(\boldsymbol x,t,\delta)$$
$$+o_{\hat P}(n^{-1/2}+(\hat{\boldsymbol\alpha}_n^*-\boldsymbol\alpha_0)). \tag{B.2.24}$$

Equality (B.2.24) follows by similar arguments used in the proof of (B.2.23) based on asymptotic equicontinuity using the closeness of $\hat F_{n,\boldsymbol\alpha}^*$ to $F_{\boldsymbol\alpha}$ and using entropy results for the functions $u\mapsto\hat F_{n,\boldsymbol\alpha}^*(u)$ and the simpler parametric functions $u\mapsto F_{\boldsymbol\alpha}(u)$ and $u\mapsto\phi_{\boldsymbol\alpha}(u)$, parametrized by the finite dimensional parameter $\boldsymbol\alpha$.

Appendix C

Single index models - Appendix

In this chapter, we give the proof of Chapter 4. Before giving the proofs, we first introduce some notations and definitions used in the remainder of this Appendix. We will denote the L_2-norm of a function f defined on $\mathcal{X} \times \mathbb{R}$ with respect to some probability measure $\mathbb{P}$ by $\|\cdot\|_{\mathbb{P}}$; i.e.

$$\|f\|_{\mathbb{P}} = \mathbb{P}(f^2)^{1/2} = \left(\int_{\mathcal{X}} f^2(\boldsymbol{x}, y) d\mathbb{P}(\boldsymbol{x}, y)\right)^{1/2}.$$

Also, we will denote by $\|\cdot\|_{B,\mathbb{P}}$ the Bernstein norm of a function f defined on $\mathcal{X} \times \mathbb{R}$ which is given by

$$\|f\|_{\mathbb{P},B} = \left(2\mathbb{P}\Big(e^{|f|} - |f| - 1\Big)\right)^{1/2}.$$

For both norms, $\mathbb{P}$ will taken to be P_0, i.e. the true joint probability measure of the $(\boldsymbol{X}, Y)$. Note that when f is only a function of $\boldsymbol{x} \in \mathcal{X}$, then

$$\|f\|_{P_0} = \left(\int_{\mathcal{X}} f^2(\boldsymbol{x}) dG(\boldsymbol{x})\right)^{1/2} = \left(\int_{\mathcal{X}} f^2(\boldsymbol{x}) g(\boldsymbol{x}) d\boldsymbol{x}\right)^{1/2},$$

by Assumption A5.

For a class of functions $\mathcal{F}$ on $\mathcal{R}$ equipped with a norm $\|\cdot\|$, we let $N_B(\zeta, \mathcal{F}, \|\cdot\|)$ denote again the minimal number N for which there exists pairs of functions $\{[g_j^L, g_j^U], 1 \leq j \leq N\}$ such that $\|g_j^U - g_j^L\| \leq \zeta$ for all $1 \leq j \leq N$ and such that for each $g \in \mathcal{F}$ there is a $j \in \{1, \ldots, N\}$ such that $g_j^L \leq g \leq g_j^U$. The $\zeta-$entropy with bracketing of $\mathcal{F}$ is defined as $H_B(\zeta, \mathcal{F}, \|\cdot\|) = \log(N_B(\zeta, \mathcal{F}, \|\cdot\|))$.

Results on entropy calculations used in proving our main results are given in Section C.4. Our proofs use inequalities for empirical processes described in Lemma 3.4.2 and Lemma 3.4.3 of van der Vaart and Wellner (1996).

Lemma 3.4.2 (van der Vaart and Wellner (1996)) Let $\mathcal{F}$ be a class of measurable functions such that $\|f\|_{\mathbb{P}} \le \delta$ and $\|f\|_\infty \le M$ for every f in $\mathcal{F}$. Then

$$\mathbb{E}\Big[\|\mathbb{G}_n\|_{\mathcal{F}}\Big] \lesssim J_n(\delta, \mathcal{F}, \|\cdot\|_{\mathbb{P}})\left(1 + \frac{J_n(\delta, \mathcal{F}, \|\cdot\|_{\mathbb{P}})}{\sqrt{n}\delta^2}M\right),$$

where

$$J_n(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + H_B(\epsilon, \mathcal{F}, \|\cdot\|)}d\epsilon$$

Lemma 3.4.3 (van der Vaart and Wellner (1996)) Let $\mathcal{F}$ be a class of measurable functions such that $\|f\|_{\mathbb{P},B} \le \delta$ for every f in $\mathcal{F}$. Then

$$\mathbb{E}\Big[\|\mathbb{G}_n\|_{\mathcal{F}}\Big] \lesssim J_n(\delta, \mathcal{F}, \|\cdot\|_{\mathbb{P},B})\left(1 + \frac{J_n(\delta, \mathcal{F}, \|\cdot\|_{\mathbb{P},B})}{\sqrt{n}\delta^2}\right),$$

In the sequel, and whenever the ϵ-bracketing entropy of some class $\mathcal{F}$ with respect to some norm $\|\cdot\|$ is bounded above by $C\epsilon^{-1}$ for some constant $C > 0$ (which may depend on n), we will write for all $e > \epsilon$

$$J_n(d) = \int_0^e (1 + C/\epsilon)^{1/2} d\epsilon. \tag{C.0.1}$$

Moreover, we will use the inequality

$$J_n(d) \le d + 2C^{1/2}e^{1/2}, \tag{C.0.2}$$

which is an immediate consequence of the fact that $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ for all $x, y \ge 0$. This Appendix is organized as follows. In Section C.1 we prove the results given in Section 4.2. The asymptotic behavior of the simple and efficient score estimator given in Section 4.3.1 and Section 4.3.2 respectively is proved in Section C.2 respectively Section C.3. The last two sections C.4 and C.5 of this Appendix contain some technical lemmas on entropy results and some auxiliary results on the behavior of the score functions. We will regularly refer to these results in the first part of this Appendix.

C.1 The least squares estimator $\hat\psi_{n,\alpha}$

In this section we first prove Proposition 4.2.1. We next show in Lemma C.1.1 that the LSE $\hat\psi_{n,\boldsymbol{\alpha}}$ is of order $O_p(\log n)$ uniformly in $\mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$. This result is used in the proof of Proposition 4.2.2, given at the end of this section.

Proof of Proposition 4.2.1. Note that with $\boldsymbol{X} \sim g$ we can write

$$L_{\boldsymbol{\alpha}}(\psi) = \mathbb{E}\Big[\Big(\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X}) - \psi(\boldsymbol{\alpha}^T\boldsymbol{X})\Big)^2\Big].$$

Thus,

$$\mathbb{E}\Big[\Big(\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X}) - \mathbb{E}(\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X}) \mid \boldsymbol{\alpha}^T\boldsymbol{X})\Big)^2\Big] = \min_{\psi\in\mathcal{M}'} L_{\boldsymbol{\alpha}}(\psi),$$

with $\mathcal{M}'$ the set of all bounded Borel-measurable function defined on $\mathcal{I}_{\boldsymbol{\alpha}} = \Big\{\boldsymbol{\alpha}^T\boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}\Big\}$. Therefore, if the minimizing function $u \mapsto \mathbb{E}(\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X}) \mid \boldsymbol{\alpha}^T\boldsymbol{X} = u)$ is monotone increasing on $\mathcal{I}_{\boldsymbol{\alpha}}$, then this implies that it necessarily minimizes $L_{\boldsymbol{\alpha}}$ over $\mathcal{M}$. Furthermore, such a minimizer is unique by strict convexity of $L_{\boldsymbol{\alpha}}$. □

Lemma C.1.1.

$$\max_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} \sup_{\boldsymbol{x}\in\mathcal{X}} \Big|\hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\Big| = O_p(\log n).$$

Proof. The proof of this lemma is similar to that of Lemma 4.4 of Balabdaoui et al. (2016). For a fixed $\boldsymbol{\alpha}$ it follows from the min-max formula of an isotonic regression that we have for all $\boldsymbol{x} \in \mathcal{X}$

$$\min_{1\le k\le n} \frac{\sum_{i=1}^k Y_i^{\boldsymbol{\alpha}}}{k} \le \hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x}) \le \max_{1\le k\le n} \frac{\sum_{i=k}^n Y_i^{\boldsymbol{\alpha}}}{n-k+1}.$$

Hence,

$$\min_{1\le i\le n} Y_i \le \hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x}) \le \max_{1\le i\le n} Y_i,$$

and this in turn implies that

$$\max_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} \sup_{\boldsymbol{x}\in\mathcal{X}} \Big|\hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\Big| \le \max_{1\le i\le n} |Y_i|.$$

Using similar arguments as in Balabdaoui et al. (2016), we use Assumption A7 to show that $\max_{1\le i\le n}|Y_i| = O_p(\log n)$, which completes the proof. □

Proof of Proposition 4.2.2. By the definition of the LSE of the unknown monotone link function, $\hat{\psi}_{n,\boldsymbol{\alpha}}$ maximizes the map $\psi \mapsto \mathbb{M}_n$ over $\mathcal{M}$ where,

$$\mathbb{M}_n(\psi, \boldsymbol{\alpha}) = \int_{\mathcal{X}\times\mathbb{R}} \Big(2y\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi^2(\boldsymbol{\alpha}^T\boldsymbol{x})\Big) d\mathbb{P}_n(\boldsymbol{x}, y). \tag{C.1.1}$$

Moreover, $\psi_{\boldsymbol{\alpha}}$ maximizes the map $\psi \mapsto \mathbb{M}$ over $\mathcal{M}$, where

$$\mathbb{M}(\psi, \boldsymbol{\alpha}) = \int_{\mathcal{X}\times\mathbb{R}} \Big(2y\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi^2(\boldsymbol{\alpha}^T\boldsymbol{x})\Big) dP_0(\boldsymbol{x}, y).$$

Define the function $f_{\psi,\boldsymbol{\alpha}}$ by

$$f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x},y) = 2y\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi^2(\boldsymbol{\alpha}^T\boldsymbol{x}),$$

Note that by definition of the LSE as the maximizer of (C.1.1), we have

$$\int_{\mathcal{X}\times\mathbb{R}} \Big(f_{\hat{\psi}_{n,\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x},y) - f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x},y)\Big)d\mathbb{P}_n(\boldsymbol{x},y) \geq 0.$$

Moreover for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)$ and $\psi \in \mathcal{M}$, we have

$$\int_{\mathcal{X}\times\mathbb{R}} \Big(f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x},y) - f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x},y)\Big)dP_0(\boldsymbol{x},y) = -d^2_{\boldsymbol{\alpha}}(\psi,\psi_{\boldsymbol{\alpha}}),$$

where, for any $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)$ and for any two elements ψ_1 and ψ_2 in $\mathcal{M}$, we define the squared distance

$$d^2_{\boldsymbol{\alpha}}(\psi_1,\psi_2) = \int_{\mathcal{X}} \Big(\psi_2(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_1(\boldsymbol{\alpha}^T\boldsymbol{x})\Big)^2 g(\boldsymbol{x})d\boldsymbol{x}.$$

This can be seen as follows:

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathbb{R}} \Big(f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x},y) - f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x},y)\Big)dP_0(\boldsymbol{x},y)\\
&= \int_{\mathcal{X}\times\mathbb{R}} \Big(2\psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})(\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})) - \psi^2(\boldsymbol{\alpha}^T\boldsymbol{x}) + \psi^2_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\Big)dP_0(\boldsymbol{x},y)\\
&= -\int_{\mathcal{X}} \Big(\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\Big)^2 g(\boldsymbol{x})d\boldsymbol{x} = -d^2_{\boldsymbol{\alpha}}(\psi,\psi_{\boldsymbol{\alpha}}),
\end{aligned}$$

where we use that $\mathbb{E}\{Y|\boldsymbol{\alpha}^T\boldsymbol{X} = u\} = \psi_{\boldsymbol{\alpha}}(u)$. This implies that, for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)$ and $\psi \in \mathcal{M}$, we have

$$\int_{\mathcal{X}\times\mathbb{R}} \Big(f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x},y) - f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x},y)\Big)d(\mathbb{P}_n - P_0)(\boldsymbol{x},y) \geq d^2_{\boldsymbol{\alpha}}(\psi,\psi_{\boldsymbol{\alpha}}).$$

We write,

$$\mathbb{P}\left\{\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} d_{\boldsymbol{\alpha}}(\hat{\psi}_{n,\boldsymbol{\alpha}},\psi_{\boldsymbol{\alpha}})\ge\epsilon\right\}$$

$$\le\mathbb{P}\left\{\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0),d_{\boldsymbol{\alpha}}(\hat{\psi}_{n,\boldsymbol{\alpha}},\psi_{\boldsymbol{\alpha}})\ge\epsilon}\left\{\int_{\mathcal{X}\times\mathbb{R}}\Big(f_{\hat{\psi}_{n,\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x},y)-f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x},y)\Big)d(\mathbb{P}_n-P_0)(\boldsymbol{x},y)\right.\right.$$
$$\left.\left.-d_{\boldsymbol{\alpha}}^2(\hat{\psi}_{n,\boldsymbol{\alpha}},\psi_{\boldsymbol{\alpha}})\right\}\ge 0,\ \sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} d_{\boldsymbol{\alpha}}(\hat{\psi}_{n,\boldsymbol{\alpha}},\psi_{\boldsymbol{\alpha}})\ge\epsilon\right\}$$

$$\le\mathbb{P}\left\{\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0),\psi\in\mathcal{M}_{RK},d_{\boldsymbol{\alpha}}(\psi,\psi_{\boldsymbol{\alpha}})\ge\epsilon}\left\{\int_{\mathcal{X}\times\mathbb{R}}\Big(f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x},y)-f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x},y)\Big)d(\mathbb{P}_n-P_0)(\boldsymbol{x},y)\right.\right.$$
$$\left.\left.-d_{\boldsymbol{\alpha}}^2(\psi_{\boldsymbol{\alpha}},\psi)\right\}\ge 0,\ \max_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right|\le K\right\}$$
$$+\,\mathbb{P}\left\{\max_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right|>K\right\}.$$

Fix $\nu>0$. Since

$$\max_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right|=O_p(\log n),$$

by Lemma C.1.1, we can find $K_1>0$ large enough such that

$$\mathbb{P}\left\{\max_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right|>K_1\log n\right\}<\nu/2.$$

Define

$$\mathcal{M}_{RK}=\Big\{\psi\text{ monotone nondecreeasing on }[-R,R]\text{ and bounded by }K\Big\},\quad\text{(C.1.2)}$$

and consider the related class

$$\mathcal{F}_{RK}=\Big\{f(\boldsymbol{x},y)=2y\Big(\psi(\boldsymbol{\alpha}^T\boldsymbol{x})-\psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\Big)-\psi(\boldsymbol{\alpha}^T\boldsymbol{x})^2+\psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})^2,$$
$$(\boldsymbol{\alpha},\psi)\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)\times\mathcal{M}_{RK}\text{ and }(\boldsymbol{x},y)\in\mathcal{X}\times\mathbb{R}\Big\},$$

and for some $v>0$

$$\mathcal{F}_{RKv}:=\Big\{f\in\mathcal{F}_{RK}:d_{\boldsymbol{\alpha}}(\psi,\psi_{\boldsymbol{\alpha}})\le v\text{ for all }\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)\Big\}.$$

Note now that the class $\mathcal{F}_{RKv}$ is included in the class $\mathcal{H}_{RC\delta}$ defined in Lemma C.4.4 given in Section C.4 with $C=2K^2$ and $\delta=2Kv$. This holds true provided that $K_0\le K$, and $K\ge 1$ which we can assume for n large enough since K will be chosen to be of order

$\log n$. To see the claimed inclusion, it is enough to show that if m is a nondecreasing function $[-R, R]$ then m^2 can be written as the difference of two monotone functions. This is true because $m^2 = m^2 \mathbb{I}_{m \geq 0} - (-m^2)\mathbb{I}_{m<0}$, and m^2 and $-m^2$ are nondecreasing on the subsets $\{m \geq 0\}$ and $\{m < 0\}$ respectively. When restricting attention to the event that $\hat{\psi}_{n,\boldsymbol{\alpha}}$ is bounded by K for n large enough, we can consider only monotone functions $\psi \in \mathcal{M}_{RK}$. Using the expression of $\psi_{\boldsymbol{\alpha}}$ the latter is bounded by $K_0 \leq K$. On the other hand, for any function $f \in \mathcal{F}_{RKv}$, there exist nondecreasing monotone functions f_1 and f_2 such that $\psi^2 - \psi_{\boldsymbol{\alpha}}^2 = f_2 - f_1$, such that $\|f_1\|_\infty, \|f_2\|_\infty \leq K^2 + K_0^2 \leq 2K^2$. Using that $K \geq 1$ implies that $\|2\psi\|_\infty, \|2\psi_{\boldsymbol{\alpha}}\|_\infty \leq 2K \leq 2K^2$. To finish, note that for any $\boldsymbol{\alpha}$ we have that $\int_{\mathcal{X}} \left(\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right)^2 dG(\boldsymbol{x}) \leq v^2$ we also have that

$$\begin{aligned}\int_{\mathcal{X}} \left(\psi^2(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}^2(\boldsymbol{\alpha}^T\boldsymbol{x})\right)^2 dG(\boldsymbol{x}) &\leq (2K)^2 \int_{\mathcal{X}} \left(\psi(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right)^2 dG(\boldsymbol{x}) \\ &\leq 4K^2v^2.\end{aligned}$$

The calculation above implies that we can take $\delta = 2Kv$. Using the result of Lemma C.4.4 in Section C.4, it follows that the related class $\widetilde{\mathcal{F}}_{RKv} = \tilde{D}^{-1}\mathcal{F}_{RKv}$ with $\tilde{D} = 16M_0C = 32M_0K^2$ and a given $v > 0$ satisfies

$$\begin{aligned} H_B\big(\epsilon, \widetilde{\mathcal{F}}_{RKv}, \|\cdot\|_{B,P_0}\big) &\leq H_B\big(\epsilon, \widetilde{\mathcal{H}}_{RKv}, \|\cdot\|_{B,P_0}\big) \\ &\leq \frac{A}{\epsilon}, \end{aligned}$$

for some constant $A > 0$ (depending only on d and the other parameters of the problem), and that for all $\tilde{f} \in \widetilde{F}_{RKv}$ we have $\|\tilde{f}\|_{B,P_0} \leq \tilde{D}^{-1}\delta = (32M_0K^2)^{-1}2Kv = (16M_0)^{-1}K^{-1}v \equiv A_0K^{-1}v$. It follows from Lemma 3.4.3 of van der Vaart and Wellner (1996) that, with J_n is defined in (C.0.1),

$$\begin{aligned} \mathbb{E}\Big[\|\mathbb{G}_n\|_{\widetilde{F}_{RKv}}\Big] &\lesssim J_n(A_0K^{-1}v), \left(1 + K^2\frac{J_n(A_0K^{-1}v)}{\sqrt{n}A_0^2v^2}\right), \\ &\leq A_0K^{-1}v + 2A_0^{1/2}K^{-1/2}v^{1/2}A^{1/2}, \quad \text{using inequality (C.0.2)} \\ &\leq B_0K^{-1/2}(v + v^{1/2}), \end{aligned}$$

for some constant $B_0 > 0$, where we used the fact that $K^{-1/2} \geq K^{-1}$. Therefore

$$\begin{aligned} \mathbb{E}\Big[\|\mathbb{G}_n\|_{\widetilde{F}_{RKv}}\Big] &\lesssim B_0K^{-1/2}(v + v^{1/2})\left(1 + K^2\frac{B_0K^{-1/2}(v + v^{1/2})}{\sqrt{n}A_0^2v^2}\right) \\ &\leq C_0K^{-1/2}(v + v^{1/2})\left(1 + C_0K^{3/2}\frac{1 + v^{1/2}}{\sqrt{n}v^{3/2}}\right). \end{aligned}$$

Using the definition of the class $\widetilde{F}_{RKv}$, the preceding display implies that

$$\mathbb{E}\Big[\|\mathbb{G}_n\|_{\mathcal{F}_{RKv}}\Big] \lesssim C_0K^{3/2}(v + v^{1/2})\left(1 + C_0K^{3/2}\frac{1 + v^{1/2}}{\sqrt{n}v^{3/2}}\right). \tag{C.1.3}$$

Now with $K = K_1 \log n$, we have that

$$\mathbb{P}\left\{\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} d_{\boldsymbol{\alpha}}(\hat{\psi}_{n,\boldsymbol{\alpha}},\psi_{\boldsymbol{\alpha}}) \geq \epsilon\right\}$$

$$\leq \mathbb{P}\left\{\sup_{\substack{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0),\psi\in\mathcal{M}_{RK},\\ d_{\boldsymbol{\alpha}}(\psi,\psi_{\boldsymbol{\alpha}})\geq\epsilon}} \left\{\int_{\mathcal{X}\times\mathbb{R}} \Big(f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x},y) - f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x},y)\Big)d(\mathbb{P}_n - P_0)(\boldsymbol{x},y)\right.\right.$$
$$\left.\left. -d^2_{\boldsymbol{\alpha}}(\psi,\psi_{\boldsymbol{\alpha}})\right\} \geq 0, \max_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right| \leq K\right\} + \nu/2$$

$$\leq \sum_{s=0}^{\infty}\mathbb{P}\left\{\sup_{\substack{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0),\psi\in\mathcal{M}_{RK},\\ 2^s\epsilon\leq d_{\boldsymbol{\alpha}}(\psi,\psi_{\boldsymbol{\alpha}})\leq 2^{s+1}\epsilon}} \sqrt{n}\int_{\mathcal{X}\times\mathbb{R}} \Big(f_{\psi,\boldsymbol{\alpha}}(\boldsymbol{x},y) - f_{\psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}}(\boldsymbol{x},y)\Big)d(\mathbb{P}_n - P_0)(\boldsymbol{x},y)\right.$$
$$\left. \geq \sqrt{n}2^{2s}\epsilon^2,\right\} + \nu/2$$

$$\leq \sum_{s=0}^{\infty}\mathbb{P}\left\{\sup_{h\in\mathcal{F}_{RK2^{s+1}\epsilon}} \sqrt{n}\int_{\mathcal{X}\times\mathbb{R}} h(\boldsymbol{x},y)d(\mathbb{P}_n - P_0)(\boldsymbol{x},y) \geq \sqrt{n}2^{2s}\epsilon^2\right\} + \nu/2. \quad \text{(C.1.4)}$$

We now show that there exists a constant $C > 0$ such that with $\epsilon = M(\log n)n^{-1/3}$,

$$\mathbb{E}\left\{\sup_{h\in\mathcal{F}_{RK2^{s+1}\epsilon}} \sqrt{n}\left|\int_{\mathcal{X}\times\mathbb{R}} h(\boldsymbol{x},y)d(\mathbb{P}_n - P_0)(\boldsymbol{x},y)\right|\right\} \leq CM^{1/2}(\log n)^2 n^{-1/6}2^{(s+1)/2}. \quad \text{(C.1.5)}$$

An application of Markov's inequality, together with (C.1.4), then yields, with $\epsilon = M(\log n)n^{-1/3}$

$$\begin{aligned}\mathbb{P}\left\{\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} d_{\boldsymbol{\alpha}}(\hat{\psi}_{n,\boldsymbol{\alpha}},\psi_{\boldsymbol{\alpha}}) \geq \epsilon\right\} &\leq \sum_{s=0}^{\infty}\frac{CM^{1/2}(\log n)^2 n^{-1/6}2^{(s+1)/2}}{\sqrt{n}2^{2s}\epsilon^2} + \nu/2\\ &= \sum_{s=0}^{\infty}\frac{C(\log n)^2 n^{-1/6}2^{(s+1)/2}}{\sqrt{n}2^{2s}M^{3/2}(\log n)^2 n^{-2/3}} + \nu/2\\ &= \frac{2^{1/2}C}{M^{3/2}}\sum_{s=0}^{\infty}\frac{1}{2^{3s/2}} + \nu/2 \leq \nu,\end{aligned}$$

for M sufficiently large. The result of Proposition 4.2.2 hence follows by showing (C.1.5). Using the obtained bound in (C.1.3) with $v = 2^{s+1}\epsilon$ and using that $2^{s+1} \geq 1, s \geq 0$ and

$\epsilon \le 1$ for n large enough we get some some constant $D_0 > 0$

$$\begin{aligned}
&\mathbb{E}\left\{\sup_{h\in\mathcal{F}_{RK2^{s+1}\epsilon}} \sqrt{n}\left|\int_{\mathcal{X}\times\mathbb{R}} h(\boldsymbol{x},y)d(\mathbb{P}_n - P_0)(\boldsymbol{x},y)\right|\right\}\\
&\lesssim (\log n)^{3/2}M^{1/2}2^{(s+1)/2}(\log n)^{1/2}n^{-1/6}\\
&\qquad\qquad \cdot\left(1 + D_0(\log n)^{3/2}\frac{1 + M^{1/2}2^{(s+1)/2}(\log n)^{1/2}n^{-1/6}}{\sqrt{n}M^{3/2}2^{3(s+1)/2}(\log n)^{3/2}n^{-1/2}}\right)\\
&= (\log n)^2 n^{-1/6}M^{1/2}2^{(s+1)/2}\left(1 + D_0'\frac{1 + M^{1/2}2^{(s+1)/2}(\log n)^{1/2}n^{-1/6}}{2^{3(s+1)/2}}\right)\\
&\le 2(\log n)^2 n^{-1/6}M^{1/2}2^{(s+1)/2},
\end{aligned}$$

with $D_0' = D_0 M^{-3/2}$, for $s \ge 0$ and n large enough. This proves the desired result:

$$\begin{aligned}
\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} d^2_{\boldsymbol{\alpha}}(\hat\psi_{n,\boldsymbol{\alpha}},\psi_{\boldsymbol{\alpha}}) &= \sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)}\int\left\{\hat\psi_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right\}^2 dG(\boldsymbol{x})\\
&= O_p\left((\log n)^2 n^{-2/3}\right).
\end{aligned}$$

□

C.2 Asymptotic behavior of the SSE

In this Section we prove Theorem 4.3.1 given in Section 4.3.1. The proof is decomposed into three parts: In Section C.2.1 we first prove the existence of a crossing of zero of $\xi_{1,n}$ defined in (4.3.5). The proof of consistency and asymptotic normality of $\hat{\boldsymbol{\alpha}}_n$ are given in Section C.2.2 and Section C.2.3.

C.2.1 Proof of existence of a crossing of zero

Let ξ be the population version of $\xi_{1,n}$ defined by

$$\xi(\boldsymbol{\beta}) \stackrel{\text{def}}{=} \int (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T\boldsymbol{x}\left\{y - \psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x},y), \tag{C.2.1}$$

where $\psi_{\boldsymbol{\alpha}}$ is defined by

$$\psi_{\boldsymbol{\alpha}}(u) \stackrel{\text{def}}{=} \mathbb{E}\left[\psi_0\left(\boldsymbol{\alpha}_0^T\boldsymbol{X}\right)|\boldsymbol{\alpha}^T\boldsymbol{X} = u\right] \equiv \mathbb{E}\left[\psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X}\right)|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\right].$$

We have the following result:

Proposition C.2.1.

$$\xi_{1,n}(\boldsymbol{\beta}) = \xi(\boldsymbol{\beta}) + o_p(1),$$

uniformly in $\boldsymbol{\beta} \in \mathcal{C} \stackrel{\text{def}}{=} \left\{\boldsymbol{\beta}\in\mathbb{R}^{d-1} : \mathbb{S}(\boldsymbol{\beta}) \in \mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)\right\}$.

Proof. For any $\boldsymbol{\beta} \in \mathcal{C}$, we write,

$$\begin{aligned}
\xi_{1,n}(\boldsymbol{\beta}) &= \int \left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T \boldsymbol{x} \left\{y - \psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\} \, d\mathbb{P}_n(\boldsymbol{x}, y) \\
&\quad + \int \left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T \boldsymbol{x} \left\{\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right) - \hat{\psi}_{n,\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\} \, d\mathbb{P}_n(\boldsymbol{x}, y) \\
&= \xi(\boldsymbol{\beta}) + \int \left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T \boldsymbol{x} \left\{y - \psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\} \, d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\
&\quad + \int \left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T \boldsymbol{x} \left\{\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right) - \hat{\psi}_{n,\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\} \, d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\
&\quad + \int \left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T \boldsymbol{x} \left\{\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right) - \hat{\psi}_{n,\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\} \, dP_0(\boldsymbol{x}, y) \\
&= \xi(\boldsymbol{\beta}) + I + II + III. \qquad\qquad \text{(C.2.2)}
\end{aligned}$$

To find the rate of convergence of the term I in (C.2.2) we will use Lemma C.4.5 in Section C.4. Note first that for $1 \le i \le d-1$ the i-th component of the vector $\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T \boldsymbol{x}$ can be written as $s(\boldsymbol{\beta})_{i1}x_1 + \ldots + s(\boldsymbol{\beta})_{id}x_d$, where by Assumption A8 the functions s_{ij} are assumed to be uniformly bounded with partial derivatives that are also uniformly bounded on the bounded convex set $\mathcal{C}$ to which $\boldsymbol{\beta}$ belongs. If B_1 is the same constant found in Lemma C.4.5 in Section C.4, then the ϵ-bracketing entropy is bounded above by B_1K_0/ϵ. Applying Lemma 3.4.3 of van der Vaart and Wellner (1996), Markov's inequality and Lemma C.4.5 in Section C.4 to each of the empirical processes corresponding to the term $s(\boldsymbol{\beta})_{ij}x_j$ for $1 \le j \le d$ yields (with $D = 8MRK_0$, and M is a constant bounding the sum of $s(\boldsymbol{\beta})_{ij}$ and their partial derivatives) for $A > 0$ and for J_n is defined in (C.0.1) that

$$\begin{aligned}
P\Big(|I| \ge An^{-1/2}\Big) &\le \frac{D}{A} J_n(B_2)\left(1 + \frac{J_n(B_2)}{\sqrt{n}B_2^2}\right) \\
&\le \frac{D}{A} B_3\left(1 + \frac{B_3}{\sqrt{n}B_2^2}\right), \quad \text{using the inequality in (C.0.2)} \\
&\asymp \frac{1}{A},
\end{aligned}$$

where B_2 is the same constant of Lemma C.4.5 and $B_3 = B_2 + 2B_1^{1/2}K_0^{1/2}B_2^{1/2}$. This implies that $I = O_p(n^{-1/2})$. For the last term III in (C.2.2) we get by an application of the Cauchy-Schwarz inequality and by Proposition 4.2.2 that this term is $O_p(n^{-1/3}\log n)$, i.e.

$$\begin{aligned}
III &\le \sqrt{\int \left\|\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T \boldsymbol{x}\right\|_2^2 dG(\boldsymbol{x})} \sqrt{\int \left\{\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right) - \hat{\psi}_{n,\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\right\}^2 dG(\boldsymbol{x})} \\
&= O_p(n^{-1/3}\log n),
\end{aligned}$$

where we also use that $\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})\right)^T \boldsymbol{x}$ is bounded in L_2 norm, a straightforward implication of Assumption A1 (boundedness of $\mathcal{X}$ and Assumption A8 (uniform boundedness of the components of the matrix $\boldsymbol{J}_{\mathbb{S}}$). The result now follows by showing that the term II is $o_p(1)$. Consider the class of functions

$$\begin{aligned}\mathcal{G}_{jRKv} &= \Big\{g(\boldsymbol{x},y) = s(\boldsymbol{\beta})x_j\left\{\psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x}) - \psi(\boldsymbol{\alpha}^T\boldsymbol{x})\right\}, \\ &\qquad \text{such that } (\boldsymbol{\alpha},\boldsymbol{\beta},\psi) \in \mathcal{S}_{d-1}\times\mathcal{C}\times\mathcal{M}_{RK} \text{ and } (\boldsymbol{x},y)\in\mathcal{X}\times\in\mathbb{R}, \\ &\qquad \text{and } \sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} d_{\boldsymbol{\alpha}}(\psi_{\boldsymbol{\alpha}},\psi)\le v\Big\},\end{aligned}$$

with s a function satisfying (C.4.3). Then, $\mathcal{G}_{jRKv}\subset\mathcal{Q}_{jRK}-\mathcal{Q}_{jRK}$, where $\mathcal{Q}_{jRC}$ is the same class defined in (C.4.4). Here, we choose K large enough such that $K\ge K_0$. If follows from (C.4.6) in the proof of Lemma C.4.5 in Section C.4, that (at the cost of increasing the constant L in (C.4.6))

$$H_B\Big(\epsilon,\tilde{\mathcal{G}}_{jRKv},\|\cdot\|_{P_0}\Big)\le\frac{LK}{\epsilon},$$

where $\tilde{\mathcal{G}}_{jRKv}=(16M_0K)^{-1}\mathcal{G}_{jRKv}$. Also, we have for all $g\in\mathcal{G}_{jRKv}$

$$\|g\|_{P_0}\le MRv.$$

Fix $\nu>0$ and let s_{ij} be the $i\times j$ entry of $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})$ for $1\le i\le d-1$ and $1\le j\le d$. Also, let

$$II_{ij}=\int_{\mathcal{X}\times\mathbb{R}}s_{ij}(\boldsymbol{\beta})x_j\Big\{\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)-\hat{\psi}_{n,\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\Big\}\,d(\mathbb{P}_n-P_0)(\boldsymbol{x},y).$$

Using Lemma C.1.1 and Proposition 4.2.2 there exists some constant $K_1>0$ large enough (and independent of n) such that with $K=K_1\log n$ and $v=K_1\log n\;n^{-1/3}$ we have that for $A>0$

$$\begin{aligned}&P\left(|II_{ij}|\ge An^{-1/2}\right)\\&=P\left(|II_{ij}|\ge An^{-1/2},\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right|\le K,\sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)}d_{\boldsymbol{\alpha}}(\psi_{\boldsymbol{\alpha}},\psi)\le v\right)\\&\quad+\nu/2\\&\lesssim\frac{K}{A}J_n(MRv)\left(1+\frac{J_n(MRv)}{\sqrt{n}M^2R^2v^2}\right)+\nu/2\\&\le\frac{K}{A}\left(MRv+2(MRL)^{1/2}K^{1/2}v^{1/2}\right)\left(1+\frac{MRv+2(MRL)^{1/2}K^{1/2}v^{1/2}}{\sqrt{n}M^2R^2v^2}\right)+\nu/2\\&\le\frac{\tilde{M}}{A}(\log n)^2\;n^{-1/6}\left(1+\frac{1}{\log nM^2R^2}\right)+\nu/2\\&\le\nu,\end{aligned}$$

for some constant $\tilde{M}>0$ and n large enough. We conclude that $II_{ij}=o_p(n^{-1/2})$ which in turn implies that $II=o_p(n^{-1/2})$. □

Proof of Theorem 4.3.1 (Existence). Using Proposition C.2.1 we get, analogously to the development in Appendix A, the relation

$$\xi_{1,n}(\boldsymbol{\alpha}) = \xi'(\boldsymbol{\beta}_0)(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + R_n(\boldsymbol{\beta}), \tag{C.2.3}$$

where $R_n(\boldsymbol{\alpha}) = o_p(1) + o(\boldsymbol{\beta}-\boldsymbol{\beta}_0)$, uniformly in $\boldsymbol{\beta}\in\mathcal{C}$ and where ξ' is the derivative of ξ defined in (C.2.1). Using Lemma C.5.1 in Section C.5, we get that the derivative of ξ at $\boldsymbol{\beta}_0$ is given by the matrix

$$\begin{aligned}\xi'(\boldsymbol{\beta}_0) &= \left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T \mathbb{E}\left[\psi_0'\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\, \mathsf{Cov}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X})\right]\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\\ &= \left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0) = \boldsymbol{B},\end{aligned}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are defined in (4.3.7) and (4.3.9) respectively. We now define, for $h>0$, the functions

$$\tilde{R}_{n,h}(\boldsymbol{\beta}) = \frac{1}{h^{d-1}}\int K_h(u_1-\beta_{01})\dots K_h(u_{d-1}-\beta_{d-1})\,R_n(u_1,\dots,u_{d-1})\,du_1\dots du_{d-1},$$

where

$$K_h(\boldsymbol{x}) = h^{-1}K(x/h), \qquad x\in\mathbb{R},$$

letting K be one of the usual smooth kernels with support $[-1,1]$.
Furthermore, we define

$$\tilde{\xi}_{1,n,h}(\boldsymbol{\beta}) = \xi'(\boldsymbol{\beta}_0)(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \tilde{R}_{nh}(\boldsymbol{\beta}).$$

Clearly

$$\lim_{h\downarrow 0}\tilde{\xi}_{1,n,h}(\boldsymbol{\beta}) = \xi_{1,n}(\boldsymbol{\beta}) \qquad \text{and} \qquad \lim_{h\downarrow 0}\tilde{R}_{nh}(\boldsymbol{\beta}) = R_n(\boldsymbol{\beta}),$$

for each continuity point $\boldsymbol{\beta}$ of $\xi_{1,n}$.
We now reparametrize, defining

$$\boldsymbol{\gamma} = \xi'(\boldsymbol{\beta}_0)\boldsymbol{\beta}, \qquad \boldsymbol{\gamma}_0 = \xi'(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0.$$

This gives

$$\xi'(\boldsymbol{\beta}_0)(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \tilde{R}_{nh}(\boldsymbol{\beta}) = \boldsymbol{\gamma}-\boldsymbol{\gamma}_0 + \tilde{R}_{nh}\left(\boldsymbol{B}^{-1}\boldsymbol{\gamma}\right),$$

By (C.2.3), the mapping

$$\boldsymbol{\gamma} \mapsto \boldsymbol{\gamma}_0 - R_n\left(\boldsymbol{B}^{-1}\boldsymbol{\gamma}\right),$$

maps, for each $\eta > 0$, the ball $B_\eta(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\} \le \eta\}$ into $B_{\eta/2}(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\} \le \eta/2\}$ for all large n, with probability tending to one, where $\|\cdot\|$ denotes the Euclidean norm, implying that the *continuous* map

$$\boldsymbol{\gamma} \mapsto \boldsymbol{\gamma}_0 - \tilde{R}_{nh}\left(\boldsymbol{B}^{-1}\boldsymbol{\gamma}\right),$$

maps $B_\eta(\boldsymbol{\gamma}_0) = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\|_2 \le \eta\}$ into itself for all large n and small h. So for large n and small h there is, by Brouwer's fixed point theorem, a point $\boldsymbol{\gamma}_{nh}$ such that

$$\boldsymbol{\gamma}_{nh} = \boldsymbol{\gamma}_0 - \tilde{R}_{nh}\left(\boldsymbol{B}^{-1}\boldsymbol{\gamma}_{nh}\right).$$

Defining $\boldsymbol{\beta}_{nh} = \boldsymbol{B}^{-1}\boldsymbol{\gamma}_{nh}$, we get

$$\tilde{\xi}_{1,n,h}(\boldsymbol{\beta}_{nh}) = \xi'(\boldsymbol{\beta}_0)(\boldsymbol{\beta}_{nh} - \boldsymbol{\beta}_0) + \tilde{R}_{nh}(\boldsymbol{\beta}_{nh}) = \mathbf{0}. \tag{C.2.4}$$

By compactness, $(\boldsymbol{\beta}_{n,1/k})_{k=1}^\infty$ must have a subsequence $(\boldsymbol{\beta}_{n,1/k_i})$ with a limit $\tilde{\boldsymbol{\beta}}_n$, as $i \to \infty$. Suppose that the jth component $\xi_{1,n,j}$ of $\xi_{1,n}$ does not have a crossing of zero at $\tilde{\boldsymbol{\beta}}_n$. Since $\xi_{1,n,j}$ only has finitely many jump discontinuities, since there can only be discontinuities at a changing of ordering of the values $\boldsymbol{\alpha}^T\boldsymbol{X}_i$, there must be a closed ball $B_\delta(\tilde{\boldsymbol{\beta}}_n) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_n\| \le \delta\}$ of $\tilde{\boldsymbol{\beta}}_n$ such that $\{\bar{\xi}_{nj}(\boldsymbol{\beta}) : \boldsymbol{\beta} \in B_\delta(\tilde{\boldsymbol{\beta}}_n)\}$ has a constant sign in the closed ball B_δ, say $\bar{\xi}_{nj}(\boldsymbol{\beta}) > 0$ for $\boldsymbol{\beta} \in \bar{B}_\delta(\tilde{\boldsymbol{\beta}}_n)$. Again using that $\xi_{1,n,j}$ only has finitely many jump discontinuities, this means that

$$\bar{\xi}_{n,j}(\boldsymbol{\beta}) \ge c > 0, \qquad \text{for all } \boldsymbol{\beta} \in \bar{B}_\delta(\tilde{\boldsymbol{\beta}}_n).$$

This means that the jth component $\tilde{\xi}_{1,n,h,j}$ of $\tilde{\xi}_{1,n,h}$ satisfies

$$\begin{aligned}
\tilde{\xi}_{1,n,h,j}(\boldsymbol{\beta}) &= \left[\xi'(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right]_j + \tilde{R}_{nh,j}(\boldsymbol{\beta})\\
&= \frac{1}{h^{d-1}}\int \left\{\left[\xi'(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right]_j + R_{nj}(u_1,\dots,u_{d-1})\right\} K_h(u_1 - \beta_{01})\\
&\qquad\qquad \dots K_h(u_{d-1} - \beta_{d-1})\,du_1\dots\,du_{d-1}\\
&\ge \frac{1}{h^{d-1}}\int \left\{\left[\xi'(\boldsymbol{\beta}_0)(\boldsymbol{u} - \boldsymbol{\beta}_0)\right]_j + R_{nj}(u_1,\dots,u_{d-1})\right\} K_h(u_1 - \beta_1)\\
&\qquad\qquad \dots K_h(u_{d-1} - \beta_{d-1})\,du_1\dots\,du_{d-1} - c/2\\
&\ge c\,\frac{1}{h^{d-1}}\int K_h(u_1 - \beta_1)\dots K_h(u_d - \beta_d)\,du_1\dots\,du_d - c/2\\
&= c/2,
\end{aligned}$$

for $\boldsymbol{\beta} \in B_{\delta/2}(\tilde{\boldsymbol{\beta}}_n)$ and sufficiently small h, contradicting (C.2.4), since $\boldsymbol{\beta}_{nh}$, for $h = 1/k_i$, belongs to $B_{\delta/2}(\tilde{\boldsymbol{\beta}}_n)$ for large k_i. □

C.2.2 Proof of consistency of the SSE

Proof of Theorem 4.3.1 (Consistency). Since $\hat{\boldsymbol{\beta}}_n$ is contained in the compact set $\mathcal{C}$, the sequence $(\hat{\boldsymbol{\beta}}_n)$ has a subsequence $(\hat{\boldsymbol{\beta}}_{n_k} = \hat{\boldsymbol{\beta}}_{n_k}(\omega))$, converging to an element $\boldsymbol{\beta}_*$. Let $\boldsymbol{\alpha}_{n_k} = \mathbb{S}(\hat{\boldsymbol{\beta}}_{n_k})$. If $\hat{\boldsymbol{\beta}}_{n_k} = \hat{\boldsymbol{\beta}}_{n_k}(\omega) \longrightarrow \boldsymbol{\beta}_*$, we get by continuity of the map $\mathbb{S}$ that $\boldsymbol{\alpha}_{n_k} \to \boldsymbol{\alpha}_* = \mathbb{S}(\boldsymbol{\beta}_*)$. By Proposition 4.2.2, we also have

$$\hat{\psi}_{n_k,\hat{\boldsymbol{\alpha}}_{n_k}}\left(\mathbb{S}(\boldsymbol{\beta}_{n_k})^T\boldsymbol{x}\right) \longrightarrow \psi_{\boldsymbol{\alpha}_*}(\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{x}),$$

where $\psi_{\boldsymbol{\alpha}}$ is defined in (4.2.3). By Proposition C.2.1 and the fact that in the limit, the crossing of zero becomes a root of the continuous limiting function, we get

$$\lim_{k\to\infty} \xi_{1,n_k}(\boldsymbol{\beta}_{n_k}) = \xi(\boldsymbol{\beta}_*) = \mathbf{0},$$

where

$$\begin{aligned}
\xi(\boldsymbol{\beta}_*) &= \int \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_*)^T\boldsymbol{x}\left\{y - \psi_{\boldsymbol{\alpha}_*}\left(\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x},y)\\
&= \int \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_*)^T\boldsymbol{x}\left\{\psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right) - \psi_{\boldsymbol{\alpha}_*}\left(\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{x}\right)\right\} dG(\boldsymbol{x})\\
&= \int \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_*)^T\boldsymbol{x}\left[\psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right) - \mathbb{E}\left\{\psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)|\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{X} = \mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{x}\right\}\right] dG(\boldsymbol{x})\\
&= \mathbb{E}\left[\mathsf{Cov}\left[\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_*)^T\boldsymbol{X}, \psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X}\right)|\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{X}\right]\right].
\end{aligned}$$

We next conclude that

$$\begin{aligned}
0 &= (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_*)^T\xi(\boldsymbol{\beta}_*)\\
&= \mathbb{E}\left[\ \mathsf{Cov}\left[(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_*)^T\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_*)^T\boldsymbol{X}, \psi_0\left(\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{X} + (\mathbb{S}(\boldsymbol{\beta}_0) - \mathbb{S}(\boldsymbol{\beta}_*))^T\boldsymbol{X}\right)|\mathbb{S}(\boldsymbol{\beta}_*)^T\boldsymbol{X}\right]\right],
\end{aligned}$$

which can only happen if $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_*$ where we use the positivity of the random variable $\mathsf{Cov}((\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T\boldsymbol{X}, \psi_0(\boldsymbol{\alpha}^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X})$ shown in Lemma C.5.2 in Section C.5 and Assumption A6 which guarantees that the random variable $\mathsf{Cov}[(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_*)^T\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})^T\boldsymbol{X}, \psi_0(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} + (\mathbb{S}(\boldsymbol{\beta}_0) - \mathbb{S}(\boldsymbol{\beta}))^T\boldsymbol{X})|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}]$ is not equal to 0 almost surely for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$. Note that

$$\begin{aligned}
&\mathsf{Cov}\left[(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^T\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})^T\boldsymbol{X}, \psi_0\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} + (\mathbb{S}(\boldsymbol{\beta}_0) - \mathbb{S}(\boldsymbol{\beta}))^T\boldsymbol{X}\right)|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\right]\\
&= \mathsf{Cov}\Big[(\mathbb{S}(\boldsymbol{\beta}_0) - \mathbb{S}(\boldsymbol{\beta}) + o(\boldsymbol{\beta} - \boldsymbol{\beta}_0))^T\boldsymbol{X}, \psi_0\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} + (\mathbb{S}(\boldsymbol{\beta}_0) - \mathbb{S}(\boldsymbol{\beta}))^T\boldsymbol{X}\right)\Big|\\
&\hspace{8cm}\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\Big]\\
&= \mathsf{Cov}\left[(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T\boldsymbol{X}, \psi_0(\boldsymbol{\alpha}^T\boldsymbol{X} + (\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = u\right] + o(\boldsymbol{\beta} - \boldsymbol{\beta}_0),
\end{aligned}$$

where the first term in the expression above is positive for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$ by Lemma C.5.2 in Section C.5.

□

C.2.3 Proof of asymptotic normality of the SSE

Proof of Theorem 4.3.1 (Asymptotic Normality). We define $\xi_{1,n}$ at $\hat{\boldsymbol{\beta}}_n$ by putting

$$\xi_{1,n}(\hat{\boldsymbol{\beta}}_n) = \mathbf{0}. \tag{C.2.5}$$

Note that, with this definition, we use the representation of the components as a convex combination of the left and right limit at $\hat{\boldsymbol{\beta}}_n$

$$\xi_{1,n,j}(\hat{\boldsymbol{\beta}}_n) = \gamma_j \xi_{1,n,j}(\hat{\boldsymbol{\beta}}_n -) + (1-\gamma_j)\xi_{1,n,j}(\hat{\boldsymbol{\beta}}_n +) = 0, \tag{C.2.6}$$

where $\xi_{1,n,j}$ denotes the jth component of $\xi_{1,n}$ and where we can choose $\gamma_j \in [0,1]$ in such a way that (C.2.6) holds since we have a crossing of zero componentwise. Note that this does not change the location of the crossing of zero. Since the following asymptotic representations are also valid for one-sided limits as used in (C.2.6) we can use Definition (C.2.5) and assume $\xi_{1,n}(\hat{\boldsymbol{\beta}}_n) = \mathbf{0}$. We show

$$\begin{aligned}
\xi_{1,n}(\hat{\boldsymbol{\beta}}_n) &= \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T \int \left\{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y - \psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\}\, d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x},y)\\
&\quad + \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T \int \left\{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\boldsymbol{\beta}_n)^T\boldsymbol{x}\right)\right\}\, dP_0(\boldsymbol{x},y)\\
&\quad + o_p\left(n^{-1/2} + \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|\right),
\end{aligned} \tag{C.2.7}$$

where from now on we will use the notation $\mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x})$ to denote $\mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = \mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x})$ for all $\boldsymbol{\beta} \in \mathcal{C}$ and $\boldsymbol{x} \in \mathcal{X}$.

Since $\hat{\boldsymbol{\beta}}_n \to_p \boldsymbol{\beta}_0$ and since the function $\boldsymbol{\beta} \to \psi_{\mathbb{S}(\boldsymbol{\beta})}(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}) \equiv \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})$ has derivative $\psi_0'(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T\big(\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X} = \mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\big)$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ for all $\boldsymbol{x} \in \mathcal{X}$ (See Lemma C.5.1 in Section C.5), we get by Definition (C.2.5) and a Taylor expansion at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ that,

$$\begin{aligned}
&\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T \int \left\{\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y - \psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\}\, d\big(\mathbb{P}_n - P_0\big)(\boldsymbol{x},y)\\
&= \boldsymbol{B}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) + o_p\left(n^{-1/2} + \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|\right),
\end{aligned}$$

where $\boldsymbol{B}$ is defined in (4.3.9). We conclude that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N_d\left(\mathbf{0}, \boldsymbol{\Pi}\right),$$

where $\boldsymbol{\Pi}$ is defined in (4.3.10). The asymptotic normality of the estimator $\hat{\boldsymbol{\alpha}}_n$ then follows by noting that

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) &= \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_p\left(\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right)\\
&\xrightarrow{d} N_d\left(\mathbf{0}, \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\boldsymbol{\Pi}\left(J_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T\right).
\end{aligned}$$

To prove (C.2.7) we first define the piecewise constant function $\bar{E}_{n,\boldsymbol{\beta}}$

$$\bar{E}_{n,\boldsymbol{\beta}}(u) = \begin{cases} \mathbb{E}\left[\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = \tau_{i,\boldsymbol{\beta}}\right] & \text{if } \psi_{\boldsymbol{\alpha}}(u) > \hat{\psi}_{n,\boldsymbol{\alpha}}(\tau_i) \;\text{ for all } u \in (\tau_i, \tau_{i+1}), \\ \mathbb{E}\left[\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = s\right] & \text{if } \psi_{\boldsymbol{\alpha}}(s) = \hat{\psi}_{n,\boldsymbol{\alpha}}(s) \;\text{ for some } s \in (\tau_i, \tau_{i+1}), \\ \mathbb{E}\left[\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = \tau_{i+1,\boldsymbol{\beta}}\right] & \text{if } \psi_{\boldsymbol{\alpha}}(u) < \hat{\psi}_{n,\boldsymbol{\alpha}}(\tau_i) \;\text{ for all } u \in (\tau_i, \tau_{i+1}), \end{cases}$$

where the $\tau_{i,\boldsymbol{\beta}}$ denote the sequence of jump points of the monotone LSE $\hat{\psi}_{n,\boldsymbol{\alpha}} = \hat{\psi}_{n,\mathbb{S}(\boldsymbol{\beta})}$. We then have

$$\int \bar{E}_{n,\hat{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\left\{y - \hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y) = \mathbf{0}. \tag{C.2.8}$$

This follows from the fact that $\hat{\psi}_{n,\boldsymbol{\alpha}}$, i.e. the minimizer of the quadratic criterion $\int_{\mathcal{X}\times\mathbb{R}}\left(y - \psi(\boldsymbol{\alpha}^T\boldsymbol{x})\right)^2 d\mathbb{P}_n(\boldsymbol{x}, y)$ over monotone functions $\psi \in \mathcal{M}$, is the left derivative of the greatest convex minorant of the cumulative sum diagram. (See also Groeneboom and Jongbloed (2014), p.332). By Lemma C.5.6 in Section C.5 we also know that $\psi'_{\boldsymbol{\alpha}}$ stays away from zero for all $\boldsymbol{\alpha} = \mathbb{S}(\boldsymbol{\beta})$ in a neighborhood of $\boldsymbol{\alpha}_0 = \mathbb{S}(\boldsymbol{\beta}_0)$. Using the same techniques as in Groeneboom and Jongbloed (2014), we can find a constant $C > 0$ such that for all $1 \le i \le d$ and $u \in \mathcal{I}_{\boldsymbol{\alpha}}$,

$$\left|\mathbb{E}\left(X_i|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\right) - \bar{E}_{ni,\boldsymbol{\beta}}(u)\right| \le C\left|\psi_{\boldsymbol{\alpha}}(u) - \hat{\psi}_{n,\boldsymbol{\alpha}}(u)\right|, \tag{C.2.9}$$

where $\bar{E}_{ni,\boldsymbol{\beta}}$ denotes the ith component of $E_{n,\boldsymbol{\beta}}$. In the sequel, we will use $\boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n) = O_p(1)$, an immediate consequence of consistency of $\hat{\boldsymbol{\alpha}}_n$ and Assumption A8. Now, as a consequence of (C.2.8), we can write

$$\begin{aligned} \xi_{1,n}(\hat{\boldsymbol{\beta}}_n) &= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\left\{y - \hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y) \\ &\quad + \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{E}_{n,\hat{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \\ &\qquad \cdot \left\{y - \hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y) \\ &= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T\left(I + II\right). \end{aligned} \tag{C.2.10}$$

The term II can be written as

$$
\begin{aligned}
II = &\int \left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{E}_{n,\hat{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \\
&\qquad\qquad \cdot \left\{y - \hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\, d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\
&+ \int \left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{E}_{n,\hat{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \\
&\qquad\qquad \cdot \left\{y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\, dP_0(\boldsymbol{x}, y) \\
&+ \int \left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{E}_{n,\hat{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \\
&\qquad\qquad \cdot \left\{\hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\, dP_0(\boldsymbol{x}, y) \\
= &\, II_a + II_b + II_c. \qquad\qquad (C.2.11)
\end{aligned}
$$

We first note that by Lemma C.5.4 in Section C.5, the functions $u \mapsto \mathbb{E}(X_i|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u)$ are uniformly bounded by R for all $\boldsymbol{\beta} \in \mathcal{C}$ and $i \in \{1, \ldots, d\}$. Also, they admit a bounded variation, with a total variation that is uniformly bounded for all $\boldsymbol{\beta} \in \mathcal{C}$ and $i \in \{1, \ldots, d\}$. By definition of $\bar{E}_{n,\boldsymbol{\beta}}$ its i-th component, $\bar{E}_{ni,\boldsymbol{\beta}}$ is also uniformly bounded by R and has a finite total variation which cannot exceed the total variation of $u \mapsto \mathbb{E}(X_i|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u)$. Using Lemma C.5.5 in Section C.5 , we can find two monotone functions f_1 and f_2 such that $u \mapsto \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\right) - \bar{E}_{n,\boldsymbol{\beta}}(u) = f_2(u) - f_1(u)$ with $f_1, f_2 \in \mathcal{M}_{RC_1}$ for some constant $C_1 > 0$. Also, we know that $\hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n} \in \mathcal{M}_{RK}$ with $K = K_1 \log n$ with increasing probability as $n \to \infty$ provided that $K_1 > 0$ is chosen large enough. Noting that for any bounded increasing functions f_1, f_2, f_3 we have that $(f_2 - f_1)f_3$ is again bounded and has a bounded variation, it follows that the class of functions, $\mathcal{F}_a$ say, involved in term II_a is included in $\mathcal{H}_{RK'v}$ defined in Lemma C.4.4 of Section C.4. Here, the constant $K' = K_2 \log n$ for some large enough constant $K_2 > 0$, and $v = C_2(\log n)^2 n^{-1/3}$ for some constant $C_2 > 0$ using (C.2.9) and Proposition 4.2.2. Using Lemma C.4.4 in Section C.4, we can show that (when the event $\hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n} \in \mathcal{M}_{RK}$ occurs)

$$
H_B\left(\epsilon, \widetilde{\mathcal{F}}_a, \|\cdot\|_{B,P_0}\right) \le \frac{B_1}{\epsilon},
$$

for some constant $B_1 > 0$, with $\widetilde{\mathcal{F}}_a = \tilde{D}^{-1}\mathcal{F}_a$ with $\tilde{D} \asymp K' = K_2 \log n$. Also, for any element $\tilde{f} = \tilde{D}^{-1} f \in \widetilde{\mathcal{F}}$ we have that

$$
\|\tilde{f}\|_{B,P_0} \le B_2 \tilde{D}^{-1} v = C_2(\log n) n^{-1/3} = \delta_n,
$$

for some constant $C_2 > 0$. Let $II_{a,i}$ be the term corresponding to i-th component of $\boldsymbol{X}$.

Using Markov's inequality we have for a fixed $A > 0$, $\nu > 0$ and n large enough that

$$
\begin{aligned}
&P\left(|II_{a,i}| \geq An^{-1/2}\right)\\
&= P\left(|II_{a,i}| \geq An^{-1/2}, \sup_{\boldsymbol{\alpha}\in\mathcal{B}(\boldsymbol{\alpha}_0,\delta_0)} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\hat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\right| \leq K\right) + \nu/2\\
&\lesssim \frac{\tilde{D}}{A} J_n(\delta_n)\left(1 + \frac{J_n(\delta_n)}{\sqrt{n}\delta_n^2}\right) + \nu/2 \lesssim \frac{\log n}{A} B_2 \delta_n^{1/2}\left(1 + \frac{B_2}{\sqrt{n}\delta_n^{3/2}}\right) + \nu/2\\
&\lesssim \frac{1}{A}(\log n)^{3/2} n^{-1/6}\left(1 + \frac{B_3}{(\log n)^{3/2}}\right) + \nu/2 \leq \nu,
\end{aligned}
$$

for some constant $B_2 > 0$ and $B_3 = B_2 C_2^{-3/2}$ using the inequality in (C.0.2) and taking n large enough. We conclude that $II_{a,i} = o_p(n^{-1/2})$ which in turn implies that

$$II_a = o_p(n^{-1/2}).$$

We turn now to II_b. Using Lemma C.5.1 in Section C.5 and a Taylor expansion of $\boldsymbol{\beta} \mapsto \psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)$ we get,

$$
\begin{aligned}
\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right) &= \psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\\
&\quad + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\left[\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T\left(\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X} = \mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right)\psi_0'\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right]\\
&\quad + o(\boldsymbol{\beta}-\boldsymbol{\beta}_0), \qquad \text{(C.2.12)}
\end{aligned}
$$

so that

$$
\begin{aligned}
II_b &= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T\int\left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{E}_{n,\hat{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\\
&\qquad\qquad \cdot\left\{\psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right) - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x},y)\\
&= o_p\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right),
\end{aligned}
$$

using consistency of $\hat{\boldsymbol{\beta}}_n$. We next consider the term II_c. Using uniform boundedness of $\boldsymbol{J}_{\mathbb{S}}$ on $\mathcal{C}$ and the inequality in (C.2.9) it follows that

$$
\begin{aligned}
\|II_c\| &\lesssim \int\left\{\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}\right) - \hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x}\right)\right\}^2 dG(\boldsymbol{x})\\
&= O_p((\log n)^2 n^{-2/3}) = o_p(n^{-1/2}),
\end{aligned}
$$

uniformly in $\beta \in \mathcal{C}$. We conclude that (C.2.10) can be written as

$$\begin{aligned}
\xi_{1,n}(\hat{\boldsymbol{\beta}}_n) &= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{ \boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \left\{ y - \hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y)\\
&\quad + o_p\left(n^{-1/2} + (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right)\\
&= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{ \boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \left\{ y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y)\\
&\quad + \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{ \boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\\
&\qquad\qquad \cdot \left\{ \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y)\\
&\quad + o_p\left(n^{-1/2} + (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right)\\
&= I_a + I_b + o_p\left(n^{-1/2} + (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right). \qquad \text{(C.2.13)}
\end{aligned}$$

We show below that $I_b = o_p\left(n^{-1/2} + (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)\right)$ such that the limiting distribution of the score estimator follows from the analysis of the term I_a which can be rewritten as

$$\begin{aligned}
I_a &= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{ \boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \left\{ y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y)\\
&\quad + \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{ \boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \left\{ y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y), \qquad \text{(C.2.14)}
\end{aligned}$$

where we recall that $\psi_{\boldsymbol{\alpha}}(u) = \mathbb{E}\left(\psi_0(\boldsymbol{\alpha}^T\boldsymbol{X}|\boldsymbol{\alpha}^T X = u\right)$. For the second term on the right-hand side of (C.2.14) we have by (C.2.12)

$$\begin{aligned}
&\boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{ \boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \left\{ y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y)\\
&= -\Bigg\{ \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T \int \psi_0'\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right) \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\}\\
&\qquad\qquad \cdot \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\}^T dP_0(\boldsymbol{x}, y)\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\Bigg\}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\\
&\quad + o_p(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0). \qquad \text{(C.2.15)}
\end{aligned}$$

For the first term on the right-hand side of (C.2.14) we have that

$$\begin{aligned}
&\boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{ \boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \left\{ y - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y)\\
&= \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\} \left\{y - \psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y)\\
&\quad + o_p(n^{-1/2}) + o_p(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0). \qquad \text{(C.2.16)}
\end{aligned}$$

Indeed, since this amounts to showing that

$$\begin{aligned}A &= \Big(J_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n) - J_{\mathbb{S}}(\boldsymbol{\beta}_0)\Big)^T \int \Big\{\boldsymbol{x} - \mathbb{E}\Big(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\Big)\Big\} \\ &\qquad \cdot \Big\{y - \psi_{\hat{\boldsymbol{\alpha}}_n}\Big(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\Big)\Big\}\, d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\ &= o_p(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0), \qquad \text{(C.2.17)} \\ B &= \int \Big(\mathbb{E}\Big(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\Big) - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\Big)(y - \psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{x}))d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\ &= o_p(n^{-1/2}), \qquad \text{(C.2.18)}\end{aligned}$$

and

$$\begin{aligned}C &= \int \Big(\boldsymbol{x} - \mathbb{E}\Big(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\Big)\Big)\Big(\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{x}) - \psi_{\hat{\boldsymbol{\alpha}}_n}\Big(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\Big)\Big)d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\ &= o_p(n^{-1/2}). \qquad \text{(C.2.19)}\end{aligned}$$

We start by proving (C.2.17). Using again that $u \mapsto \mathbb{E}\Big(X_i|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X} = u\Big)$ is a bounded function with a uniformly bounded total variation, and that x_i is a fixed (and deterministic) function, we can show that the class of functions involved in A, $\mathcal{F}_A$ say, satisfies $\mathcal{F}_A \subset x_i\mathcal{H}_{RC_1v} + \mathcal{H}_{RC_1v}$ with v and C_1 are some constants that are independent of n (since $\psi_{\boldsymbol{\alpha}}$, $\boldsymbol{X}$ and $u \mapsto E[\boldsymbol{X}|\boldsymbol{\alpha}^T X = u]$ are all bounded by constants independent of n). Now it follows by Lemma C.4.4 in Section C.4 that $H_B\big(\epsilon, \widetilde{\mathcal{H}}_{RC_1v}, \|\cdot\|_{B,P_0}\big) \lesssim 1/\epsilon$ with $\widetilde{\mathcal{H}}_{RC_1v} = (16M_0C_1)^{-1}\mathcal{H}_{RC_1v}$ and $\|\tilde{h}\|_{B,P_0} \lesssim C_2$ for some constant $C_2 > 0$ that is independent of n for all $\tilde{h} \in \widetilde{\mathcal{H}}_{RC_1v}$. Hence, using arguments similar to those of the proof of $II_a = o_p(n^{-1/2})$ we can show that

$$\int \Big\{\boldsymbol{x} - \mathbb{E}\Big(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\Big)\Big\}\Big\{y - \psi_{\hat{\boldsymbol{\alpha}}_n}\Big(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\Big)\Big\}\, d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) = O_p(n^{-1/2}).$$

Using a Taylor expansion of $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_0$ gives the desired rate in (C.2.17).

Now we turn to term B in (C.2.18). Fix $\nu > 0$ and $i \in \{1, \dots, d\}$. Using consistency of $\hat{\boldsymbol{\beta}}_n$ and Lemma C.5.3 in Section C.5, then for all $\eta > 0$ there exists n large enough such that

$$\Big|\mathbb{E}\Big(X_i|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\Big) - \mathbb{E}\left(X_i|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\Big| \le \eta,$$

with probability at least $1 - \nu/2$. Thus, for $L > 0$ we have that for the ith component

B_i of B defined in (C.2.18) and for n large enough

$$
\begin{aligned}
&P(|B_i| > Ln^{-1/2})\\
&= P\Big(\Big|\int \Big(\mathbb{E}\left(X_i|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \mathbb{E}\left(X_i|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\Big)(y - \psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{x}))d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y)\Big| \\
&\qquad\qquad > Ln^{-1/2}\Big)\\
&\le \nu/2 + \frac{1}{L}E[\|\mathbb{G}_n\|_{\mathcal{F}'}]\\
&\le \nu/2 + \frac{C_1}{L}\eta,
\end{aligned}
$$

where $\mathcal{F}'$ is defined in (C.4.9) and where we have used the result of Lemma C.4.7 for some constant $C_1 > 0$. Choosing η such that $\eta \le \nu L C_1^{-1}/2$ gives the claimed rate of convergence in (C.2.18).

To establish the convergence rate of C, we first note that, for $i \in \{1, \ldots, d\}$, we have that $\boldsymbol{x} \mapsto \mathbb{E}(X_i|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X} = \mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x})\psi_{\hat{\boldsymbol{\alpha}}_n}(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x})$ belongs to the class $\mathcal{G}_{RC_1} - \mathcal{G}_{RC_1}$ for some constant $C_1 > 0$ where $\mathcal{G}_{RK}$ was defined in (C.4.1). This follows from using again the fact that the function $u \mapsto E\left(X_i|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\right)$ is uniformly bounded and has a uniform total variation for all $\beta \in \mathcal{C}$, that $\psi_{\boldsymbol{\alpha}}$ is a bounded monotone function, and the fact that $(f_1 - f_2)f_3$ is a bounded function with bounded total variation for any increasing and bounded functions f_1, f_2 and f_3, where we again use Lemma C.5.5 in Section C.5 to write the function $u \mapsto E\left(X_i|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\right)$ as the difference $f_1 - f_2$. Note now that both $\boldsymbol{x} \mapsto x_i$ and $\boldsymbol{x} \mapsto \psi_0(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})$ are fixed and bounded functions, and that the order bracketing entropy of a class does not get altered after multiplication its members by such functions (similarly for addition). It follows from Lemma C.4.2 and Lemma C.4.3 in Section C.4, that the ϵ-bracketing entropy of the class of functions involved in term C with respect to $\|\cdot\|_{P_0}$ is bounded above by B/ϵ for some constant B.

Furthermore, using consistency of $\hat{\boldsymbol{\alpha}}_n$ and Lemma C.5.3 of Section C.5, we can find for any fixed $\nu > 0$ an $\eta > 0$ such that $\sup_{\boldsymbol{x}} |\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{x}) - \psi_{\hat{\boldsymbol{\alpha}}_n}(\hat{\boldsymbol{\alpha}}_n^T\boldsymbol{x})| \le \eta$ with probability at least $1 - \nu/2$ for n large enough. Hence, at the cost of increasing the constant B, both the $\|\cdot\|_\infty$ and $\|\cdot\|_{P_0}$ norms of the functions of the class involved in term C are bounded above by $B\eta$. Using Markov's inequality and Lemma 3.4.2 of van der Vaart and Wellner (1996) it follows that for all $L > 0$

$$
\begin{aligned}
&P(|C_i| > Ln^{-1/2})\\
&= P\Big(\Big|\int \left(x_i - \mathbb{E}\left(\boldsymbol{X}_i|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right)\left(\psi_0(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}) - \psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right)d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y)\Big|\\
&\qquad\qquad \ge Ln^{-1/2}\Big)\\
&\le \nu/2 + \frac{1}{L}J_n(B\eta)\left(1 + B\eta\frac{J_n(B\eta)}{\sqrt{n}B^2\eta^2}\right) \le \nu/2 + \frac{1}{L}\left(B_1\eta^{1/2} + \frac{B_1}{B}\frac{1}{\sqrt{n}}\right) \le \nu,
\end{aligned}
$$

taking η small enough and n large enough. We conclude that $C = o_p(n^{-1/2})$. Now we come back to term I_b given by

$$I_b = \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \cdot \left\{\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y).$$

Note first that

$$\begin{aligned} I_b &= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \\ &\qquad \cdot \left\{\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\ &= \boldsymbol{J}_{\mathbb{S}}(\hat{\boldsymbol{\beta}}_n)^T I_b', \end{aligned} \tag{C.2.20}$$

since

$$\begin{aligned} &\int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \left\{\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y) \\ &= \mathbb{E}\left[\left(\boldsymbol{X} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right)\right)\left(\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right) - \hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right)\right)\right] \\ &= \mathbb{E}\Big[\mathbb{E}\left[\left(\boldsymbol{X} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right)\right)|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right] \\ &\qquad \cdot \left(\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right) - \hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{X}\right)\right)\Big] \\ &= \mathbf{0}. \end{aligned}$$

Let $\mathcal{F}_b$ denote the class of functions involved in term I_b' defined in (C.2.20), where in the definition of this class we consider the event where $\hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}$ is bounded. Given the arguments used recurrently above we can directly state that the ϵ-bracketing entropy of this class is no larger than $A_1 \log n/\epsilon$ for some constant $A_1 > 0$ with increasing probability. Also, the $\|\cdot\|_\infty$ and $\|\cdot\|_{P_0}$ norms of the members of the class $\mathcal{F}_b$ are respectively bounded above with increasing probability by $A_1 \log n$ and $A_1 \log n \; n^{-1/3} = \eta_n$ at the cost of taking a larger A_1. For a fixed $\nu > 0$ and $L > 0$ we have for $i \in \{1, \ldots, d\}$,using Lemma 3.4.2 of van der Vaart and Wellner (1996),

$$P\Big(\Big|\int\Big\{\boldsymbol{x}_i-\mathbb{E}\left(\boldsymbol{X}_i|\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\Big\}\Big\{\psi_{\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)-\hat{\psi}_{n,\hat{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\hat{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\Big\}$$
$$d(\mathbb{P}_n-P_0)(\boldsymbol{x},y)\Big|>Ln^{-1/2}\Big)$$
$$\le \nu/2+\frac{A_2}{L}(\log n)^{1/2}\eta_n^{1/2}\left(1+\frac{A_2(\log n)^{1/2}\eta_n^{1/2}}{\sqrt{n}\eta_n^2}(\log n)\right)$$
$$\le \nu/2+\frac{A_2}{L}(\log n)^{1/2}\eta_n^{1/2}\left(1+\frac{A_2(\log n)^{3/2}}{\sqrt{n}\eta_n^{3/2}}\right)$$
$$\lesssim \nu/2+\frac{A_2}{L}(\log n)n^{-1/6}\left(1+\frac{A_2}{A_1^{3/2}}\right)\le\nu,$$

for some constant $A_2>0$ and n large enough. This implies that $I_b=o_p(n^{-1/2})$. We conclude by (C.2.15), (C.2.16), (C.2.17), (C.2.18), (C.2.19) and Definition (C.2.5) that,

$$\boldsymbol{B}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\right)=\int(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\left\{\boldsymbol{x}-\mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y-\psi_0\big(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\big)\right\}$$
$$d\big(\mathbb{P}_n-P_0\big)(\boldsymbol{x},y)$$
$$+o_p\left(n^{-1/2}+\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|\right),$$

where

$$\boldsymbol{B}=(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\,\mathbb{E}\Big[\psi_0'(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X})\ \mathsf{Cov}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X})\Big]\,(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)).$$

We get,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\right)=\sqrt{n}\boldsymbol{B}^{-1}\int(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\left\{\boldsymbol{x}-\mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right\}\left\{y-\psi_0\big(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\big)\right\}$$
$$d\big(\mathbb{P}_n-P_0\big)(\boldsymbol{x},y)$$
$$+o_p\left(1+\sqrt{n}\|\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|\right)$$
$$\xrightarrow{d} N(\mathbf{0},\boldsymbol{\Pi}),$$

where

$$\boldsymbol{\Pi}=\boldsymbol{B}^{-1}\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T\boldsymbol{\Sigma}\,\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\,\boldsymbol{B}^{-1}\in\mathbb{R}^{(d-1)\times(d-1)}.$$

The asymptotic limiting distribution of the single index score estimator $\hat{\boldsymbol{\alpha}}_n$ now follows by an application of the Delta method and we conclude that

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)=J_{\mathbb{S}}(\boldsymbol{\beta}_0)\sqrt{n}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)+o_p\left(\sqrt{n}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0)\right)$$
$$\xrightarrow{d} N_d\left(\mathbf{0},J_{\mathbb{S}}(\boldsymbol{\beta}_0)\boldsymbol{\Pi}\left(J_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T\right).$$

Finally, the result of Theorem 4.3.1 follows by Lemma 4.3.1. This completes the proof. □

C.3 Asymptotic behavior of the ESE

In this section we prove (iii) of Theorem 4.3.2 on the asymptotic normality of the efficient score estimator $\tilde{\boldsymbol{\alpha}}_n$. The proofs of existence and consistency of $\tilde{\boldsymbol{\alpha}}_n$, given in (i) and (ii) of Theorem 4.3.2 follow the same lines as the corresponding proofs for the simple score estimator $\hat{\boldsymbol{\alpha}}_n$ given in Sections C.2.1 and C.2.1 and are omitted.

Proof of Theorem 4.3.2 (Asymptotic Normality). Let τ_i denote the sequence of jump points of the monotone LSE $\hat{\psi}_{n,\boldsymbol{\alpha}}$. We introduce the piecewise constant function $\bar{\rho}_{n,\boldsymbol{\beta}}$ defined for $u \in [\tau_i, \tau_{i+1})$ as

$$\bar{\rho}_{n,\boldsymbol{\beta}}(u) =$$
$$\begin{cases} \mathbb{E}[\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = \tau_i]\psi'_{\boldsymbol{\alpha}}(\tau_i) & \text{if } \psi_{\boldsymbol{\alpha}}(u) > \hat{\psi}_{n,\boldsymbol{\alpha}}(\tau_i) \text{ for all } u \in (\tau_i, \tau_{i+1}), \\ \mathbb{E}[\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = s]\psi'_{\boldsymbol{\alpha}}(s) & \text{if } \psi_{\boldsymbol{\alpha}}(s) = \hat{\psi}_{n,\boldsymbol{\alpha}}(s) \text{ for some } s \in (\tau_i, \tau_{i+1}), \\ \mathbb{E}[\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = \tau_{i+1}]\psi'_{\boldsymbol{\alpha}}(\tau_{i+1}) & \text{if } \psi_{\boldsymbol{\alpha}}(u) < \hat{\psi}_{n,\boldsymbol{\alpha}}(\tau_i) \text{ for all } u \in (\tau_i, \tau_{i+1}). \end{cases}$$

We can write,

$$\begin{aligned} \xi_{2,nh}(\tilde{\boldsymbol{\beta}}_n) &= \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x}\tilde{\psi}'_{nh,\boldsymbol{\alpha}}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \\ &\qquad \cdot \left\{y - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y) \\ &\quad + \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int \left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{\rho}_{n,\tilde{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \\ &\qquad \cdot \left\{y - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y) \\ &= J + JJ, \end{aligned} \tag{C.3.1}$$

using,

$$\int \bar{\rho}_{n,\tilde{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\left\{y - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y) = \boldsymbol{0}.$$

The term JJ can be written as

$$\begin{aligned} JJ &= \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int \left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{\rho}_{n,\tilde{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \\ &\qquad \cdot \left\{y - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\ &\quad + \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int \left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{\rho}_{n,\tilde{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \\ &\qquad \cdot \left\{y - \psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y) \\ &\quad + \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int \left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{\rho}_{n,\tilde{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \\ &\qquad \cdot \left\{\psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y) \\ &= JJ_a + JJ_b + JJ_c, \end{aligned} \tag{C.3.2}$$

We first note that by Assumption A10, the functions $u \mapsto \psi'_{\boldsymbol{\alpha}}(u) = \psi'_{\mathbb{S}(\boldsymbol{\beta})}(u)$ are uniformly bounded and have a total variation that is uniformly bounded for all $\boldsymbol{\beta} \in \mathcal{C}$. This also implies, using Lemma C.5.4 in Section C.5, that the functions $u \mapsto \mathbb{E}\left(X_i|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\right)\psi'_{\boldsymbol{\alpha}}(u)$ have a bounded variation for all $\boldsymbol{\beta} \in \mathcal{C}$. Using the same arguments as those for term II_a defined in (C.2.11) in the proof of Theorem 4.3.1, it easily follows that,

$$JJ_a = o_p(n^{-1/2}).$$

We next consider the term JJ_b. By Lemma C.5.6 in Section C.5, we know that $\psi'_{\boldsymbol{\alpha}}$ stays away from zero for all $\mathbb{S}(\boldsymbol{\beta})$ in a neighborhood of $\mathbb{S}(\boldsymbol{\beta}_0)$. Using the same techniques as in Groeneboom and Jongbloed (2014), we can find a constant $K > 0$ such that for all $1 \le i \le d$ and $u \in \mathcal{I}_{\boldsymbol{\alpha}}$,

$$\left|\mathbb{E}\left(X_i|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\right)\psi'_{\boldsymbol{\alpha}}(u) - \bar{\rho}_{ni,\boldsymbol{\beta}}(u)\right| \le K\left|\psi_{\boldsymbol{\alpha}}(u) - \hat{\psi}_{n,\boldsymbol{\alpha}}(u)\right|, \tag{C.3.3}$$

where $\bar{\rho}_{ni,\boldsymbol{\beta}}$ denotes the ith component of $\rho_{n,\boldsymbol{\beta}}$. This implies that the difference $\mathbb{E}\left(X_i|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u\right)\psi'_{\boldsymbol{\alpha}}(u) - \bar{\rho}_{ni,\boldsymbol{\beta}}(u)$ converges to zero for all $u \in \mathcal{I}_{\boldsymbol{\alpha}}$. Recall from (C.2.12) that, using Lemma C.5.1 in Section C.5 and a Taylor expansion of $\boldsymbol{\beta} \mapsto \psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)$ we get

$$\begin{aligned}
\psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right) &= \psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\\
&\quad + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\left[\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T\left(\boldsymbol{x} - \mathbb{E}(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{X} = \mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x})\right)\psi'_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right]\\
&\quad + o(\boldsymbol{\beta}-\boldsymbol{\beta}_0),
\end{aligned}$$

such that

$$\begin{aligned}
JJ_b = \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int &\left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{\rho}_{n,\tilde{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\\
&\cdot\left\{\psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right) - \psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x},y)\\
= o_p\left(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right).&
\end{aligned}$$

For the therm JJ_c, we get by an application of the Cauchy-Schwarz inequality together with the uniform boundedness of $\boldsymbol{J}_{\mathbb{S}}$, Proposition 4.2.2 and (C.3.3) that,

$$\begin{aligned}
JJ_c &\le \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T\sqrt{\int\left\{\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \bar{\rho}_{n,\tilde{\boldsymbol{\beta}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}^2 dP_0(\boldsymbol{x},y)}\\
&\qquad\cdot\sqrt{\int\left\{\psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}^2 dP_0(\boldsymbol{x},y)}\\
&\lesssim \int\left\{\psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\tilde{\boldsymbol{\alpha}}_n^T\boldsymbol{x}\right) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}\left(\tilde{\boldsymbol{\alpha}}_n^T\boldsymbol{x}\right)\right\}^2 dG(\boldsymbol{x})\\
&= O_p\left((\log n)^2 n^{-2/3}\right) = o_p(n^{-1/2}).
\end{aligned}$$

We conclude that (C.3.1) can be written as

$$
\begin{aligned}
&\xi_{2,nh}(\tilde{\boldsymbol{\beta}}_n)\\
&= \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x}\tilde{\psi}'_{nh,\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\\
&\qquad\qquad \cdot \left\{y - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y)\\
&\quad + o_p\left(n^{-1/2} + (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right)\\
&= \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x}\tilde{\psi}'_{nh,\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\\
&\qquad\qquad \cdot \left\{y - \psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y)\\
&\quad + \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x}\tilde{\psi}'_{nh,\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\\
&\qquad\qquad \cdot \left\{\psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y)\\
&\quad + \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x}\tilde{\psi}'_{nh,\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\\
&\qquad\qquad \cdot \left\{\psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y)\\
&\quad + o_p\left(n^{-1/2} + (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right)\\
&= J_a + J_b + J_c + o_p\left(n^{-1/2} + (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right). \qquad \text{(C.3.4)}
\end{aligned}
$$

We first consider the term J_b. By Assumption A10, Lemma C.5.4 and Lemma C.5.7 in Section C.5, we get that the functions $u \mapsto \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x} = u\right)\psi'_{\tilde{\boldsymbol{\alpha}}_n}(u)$ and $u \mapsto \tilde{\psi}'_{nh,\tilde{\boldsymbol{\alpha}}_n}(u)$ have a uniformly bounded total variation for all $\boldsymbol{\beta} \in \mathcal{C}$. Using similar arguments as for the term I_b defined in (C.2.13) we get for $A > 0$ and $\nu > 0$ that

$$P(|J_b| \geq An^{-1/2}) \leq \nu,$$

for n large enough and we conclude that $J_b = o_p(n^{-1/2})$. For the term J_c we get,

$$
\begin{aligned}
J_c &= \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\tilde{\psi}'_{nh,\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\\
&\qquad\qquad \cdot \left\{\psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y)\\
&\quad + \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int \left\{\tilde{\psi}'_{nh,\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\\
&\qquad\qquad \cdot \left\{\psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y)\\
&= \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int \left\{\tilde{\psi}'_{nh,\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\\
&\qquad\qquad \cdot \left\{\psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y),
\end{aligned}
$$

Furthermore, let $H_{\boldsymbol{\beta}}$ be the distribution function of the random variable $\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}$ and let $\mathbb{E}(\boldsymbol{X}|u)$ denote the conditional expectation of $\boldsymbol{X}$ given $\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X} = u$, then

$$
\begin{aligned}
&\int \left\{\tilde{\psi}'_{nh,\tilde{\boldsymbol{\alpha}}_n}(u) - \psi'_{\tilde{\boldsymbol{\alpha}}_n}(u)\right\} \mathbb{E}\left(\boldsymbol{X}|u\right) \left\{\psi_{\tilde{\boldsymbol{\alpha}}_n}(u) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}(u)\right\} dH_{\tilde{\boldsymbol{\beta}}_n}(u) \\
&= \int \left\{\frac{1}{h}\int K\left(\{u-v\}/h\right) d\hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}(v) - \psi'_{\tilde{\boldsymbol{\alpha}}_n}(u)\right\} \\
&\qquad\qquad \cdot \mathbb{E}\left(\boldsymbol{X}|u\right) \left\{\psi_{\tilde{\boldsymbol{\alpha}}_n}(u) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}(u)\right\} dH_{\tilde{\boldsymbol{\beta}}_n}(u) \\
&= \int \left(\frac{1}{h^2}\int K'\left(\{u-v\}/h\right) \left\{\hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}(v) - \psi_{\tilde{\boldsymbol{\alpha}}_n}(v)\right\} dv\right) \\
&\qquad\qquad \cdot \mathbb{E}\left(\boldsymbol{X}|u\right) \left\{\psi_{\tilde{\boldsymbol{\alpha}}_n}(u) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}(u)\right\} dH_{\tilde{\boldsymbol{\beta}}_n}(u) \\
&\quad + \int \left(\frac{1}{h}\int K\left(\{u-v\}/h\right) \psi'_{\tilde{\boldsymbol{\alpha}}_n}(v)\, dv - \psi'_{\tilde{\boldsymbol{\alpha}}_n}(u)\right) \\
&\qquad\qquad \cdot \mathbb{E}\left(\boldsymbol{X}|u\right) \left\{\psi_{\tilde{\boldsymbol{\alpha}}_n}(u) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}(u)\right\} dH_{\tilde{\boldsymbol{\beta}}_n}(u).
\end{aligned}
$$

The last term on the right hand side is $O_p\left(n^{-2/7-1/3}\right) = o_p\left(n^{-1/2}\right)$. This follows by an application of the Cauchy-Schwarz inequality since

$$
\left\{\int \left(\frac{1}{h}\int K\left(\{u-v\}/h\right) \psi'_{\tilde{\boldsymbol{\alpha}}_n}(v)\, dv - \psi'_{\tilde{\boldsymbol{\alpha}}_n}(u)\right)^2 dH_{\tilde{\boldsymbol{\beta}}_n}(u)\right\}^{1/2} = O_p\left(n^{-2/7}\right),
$$

and

$$
\left\{\int \left(\psi_{\tilde{\boldsymbol{\alpha}}_n}(u) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}(u)\right)^2 dH_{\tilde{\boldsymbol{\beta}}_n}(u)\right\}^{1/2} = O_p\left(n^{-1/3}\right).
$$

The first term on the right hand side is $O_p\left(n^{1/7-2/3}\right) = o_p\left(n^{-1/2}\right)$ using that for small h

$$
\begin{aligned}
&\int \left(\frac{1}{h^2}\int K'\left(\{u-v\}/h\right) \left\{\hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}(v) - \psi_{\tilde{\boldsymbol{\alpha}}_n}(v)\right\} dv\right) \\
&\qquad\qquad \cdot \mathbb{E}\left(\boldsymbol{X}|u\right) \left\{\psi_{\tilde{\boldsymbol{\alpha}}_n}(u) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}(u)\right\} dH_{\tilde{\boldsymbol{\beta}}_n}(u) \\
&\lesssim \frac{1}{h}\int \left(\psi_{\tilde{\boldsymbol{\alpha}}_n}(u) - \hat{\psi}_{n\tilde{\boldsymbol{\alpha}}_n}(u)\right)^2 dH_{\tilde{\boldsymbol{\beta}}_n}(u).
\end{aligned}
$$

We conclude that (C.3.4) can be written as

$$\begin{aligned}
\xi_{2,nh}(\tilde{\boldsymbol{\beta}}_n) = \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int &\left\{\boldsymbol{x}\tilde{\psi}'_{nh,\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \\
&\cdot\left\{y - \psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d\mathbb{P}_n(\boldsymbol{x}, y) \\
&+ o_p\left(n^{-1/2} + (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right) \\
= \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int &\boldsymbol{x}\left\{\tilde{\psi}'_{nh,\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \\
&\cdot\left\{y - \psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\
+ \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int &\boldsymbol{x}\left\{\tilde{\psi}'_{nh,\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) - \psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} \\
&\cdot\left\{y - \psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y) \\
+ \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int &\left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) \\
&\cdot\left\{y - \psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\
+ \boldsymbol{J}_{\mathbb{S}}(\tilde{\boldsymbol{\beta}}_n)^T \int &\left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\}\psi'_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right) \\
&\cdot\left\{y - \psi_{\tilde{\boldsymbol{\alpha}}_n}\left(\mathbb{S}(\tilde{\boldsymbol{\beta}}_n)^T\boldsymbol{x}\right)\right\} dP_0(\boldsymbol{x}, y) \\
+ o_p&\left(n^{-1/2} + (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right) \\
= JJJ_a &+ JJJ_b + JJJ_c + JJJ_d + o_p\left(n^{-1/2} + (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\right).
\end{aligned}$$

We consider JJJ_a first and note that by Assumption A10 and Lemma C.5.7 i Section C.5, the functions $\psi'_{\boldsymbol{\alpha}}$ and $\tilde{\psi}'_{nh,\boldsymbol{\alpha}}$ have a uniformly bounded total variation. By an application of Lemma C.5.5 we can write the difference $\tilde{\psi}'_{nh,\boldsymbol{\alpha}} - \psi'_{\boldsymbol{\alpha}}$ as the difference of two monotone functions, say $f_1, f_2 \in \mathcal{M}_{RC_1}$ for some constant $C_1 > 0$. This implies that the class of functions

$$\mathcal{F}_1 = \Big\{ f(\boldsymbol{x}, y) \stackrel{\text{def}}{=} \{\tilde{\psi}'_{nh,\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right) - \psi'_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\}\{y - \psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)\}, \quad (\boldsymbol{x}, y, \boldsymbol{\beta}) \in \mathcal{X} \times \mathbb{R} \times \mathcal{C} \Big\},$$

is contained in the class $\mathcal{H}_{RC_1v}$ where $v \asymp h^{-1}\log n n^{-1/3}$ (See the proof of Lemma C.5.7 in Section C.5). By Lemma C.4.4 in Section C.4 and the fact that the order bracketing entropy of a class does not get altered after multiplication with the fixed and bounded function $\boldsymbol{x} \mapsto x_i$ we get that the class of functions involved with the term JJJ_a, say $\mathcal{F}_a$, satisfies

$$H_B\big(\epsilon, \mathcal{F}_a, \|\cdot\|_{B,P_0}\big) \lesssim \frac{1}{\epsilon} \qquad \text{and} \qquad \|f\|_{B,P_0} \lesssim v.$$

Using again an application of Markov's inequality, together with Lemma 3.4.3 of van der Vaart and Wellner (1996) we conclude that for $A > 0$

$$P(|JJJ_a| > An^{-1/2}) \lesssim v^{1/2} = h^{-1/2}(\log n)^{1/2}n^{-1/6},$$

which can be made arbitrarily small for n large enough and $h \asymp n^{-1/7}$. We conclude that

$$JJJ_a = o_p(n^{-1/2}).$$

Using similar arguments as for the term JJ_b defined in (C.3.2) we also get

$$JJJ_b = o_p\left(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right).$$

The result of Theorem 4.3.2 follows by noting that, using the same techniques as for the term I_a in (C.2.16), we get

$$\begin{aligned} JJJ_c = (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \int & \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\} \psi_0'\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right) \\ & \cdot \left\{y - \psi_0\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\} d(\mathbb{P}_n - P_0)(\boldsymbol{x}, y) \\ & + o_p(n^{-1/2}) + o_p(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0), \end{aligned}$$

and that by a Taylor expansion of $\boldsymbol{\beta} \mapsto \psi_{\boldsymbol{\alpha}}\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right)$ we get

$$\begin{aligned} JJJ_d = -\Bigg\{ & (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \left(\int \left(\psi_0'\left(\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right)^2 \cdot \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\} \right. \\ & \left. \cdot \left\{\boldsymbol{x} - \mathbb{E}\left(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta}_0)^T\boldsymbol{x}\right)\right\}^T dP_0(\boldsymbol{x}, y) \right) J_{\mathbb{S}}(\boldsymbol{\beta}_0) \Bigg\} (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\ & + o_p(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0). \end{aligned}$$

The rest of the proof follows the same line as the proof of asymptotic normality of the simple score estimator defined in Theorem 4.3.1 and is omitted. □

C.4 Entropy results

Lemma C.4.1. *Fix $\epsilon > 0$, and consider $\mathcal{F}_1$ a class of functions defined on $\mathcal{X} \times \mathbb{R}$ bounded by some constant $A > 0$ and equipped by the L_2 norm $\|\cdot\|_{P_0}$ with respect to P_0. Also, let $\mathcal{F}_2$ be another class of continuous functions defined on a bounded set $\mathcal{C} \subset \mathbb{R}^{d-1}$ such that $\mathcal{F}_2$ is equipped by the supremum norm $\|\cdot\|_\infty$, and bounded by some constant $B > 0$. Moreover assume that $H_B(\epsilon, \mathcal{F}_1, \|\cdot\|_{P_0}) < \infty$ and $H_B(\epsilon, \mathcal{F}_2, \|\cdot\|_\infty) < \infty$. Consider*

$$\mathcal{F} = \mathcal{F}_1\mathcal{F}_2 = \Big\{f(\boldsymbol{x}) = f_{\boldsymbol{\beta}}(\boldsymbol{x}, y) = f_1(\boldsymbol{x}, y)f_2(\boldsymbol{\beta}) : (\boldsymbol{x}, y, \boldsymbol{\beta}) \in \mathcal{X} \times \mathbb{R} \times \mathcal{C}\Big\}.$$

Then there exists some constant $B > 0$ such that

$$H_B(\epsilon, \mathcal{F}, \|\cdot\|_{P_0}) \le H_B(B\epsilon, \mathcal{F}_1, \|\cdot\|_{P_0}) + H_B(B\epsilon, \mathcal{F}_2, \|\cdot\|_\infty).$$

Proof. Let $f = f_1 f_2 \in \mathcal{F}$ for some pair $(f_1, f_2) \in \mathcal{F}_1 \times \mathcal{F}_2$. For $\epsilon > 0$ consider the (f_1^L, f_1^U) and (f_2^L, f_2^U) ϵ-brackets with respect to $\|\cdot\|_{P_0}$ for f_1 and f_2. Note that since $\mathcal{F}_1$ and $\mathcal{F}_2$ are bounded by $M = \max(A, B)$ we can always assume that $-M \leq f_i^L \leq f_i^U \leq M$ for $i \in \{1, 2\}$. As we deal with a product of two functions, construction of a bracket for f requires considering different sign cases for a given pair $(\boldsymbol{x}, \boldsymbol{\beta})$:

1. $0 \leq f_1^L(\boldsymbol{x})$ and $0 \leq f_2^L(\boldsymbol{\beta})$,
2. $0 \leq f_1^L(\boldsymbol{x})$, $f_2^L(\boldsymbol{\beta}) < 0$ and $f_2^U(\boldsymbol{\beta}) \geq 0$,
3. $f_1^L(\boldsymbol{x}) \leq 0$, $f_1^U(\boldsymbol{x}) \geq 0$ and $0 \leq f_2^L(\boldsymbol{\beta})$,
4. $f^U(\boldsymbol{x}) \leq 0$, $f^L(\boldsymbol{\beta}) \geq 0$,
5. $f^L(\boldsymbol{x}) \geq 0$, $f^U(\boldsymbol{\beta}) \geq 0$,
6. $f_1^L(\boldsymbol{x}) \leq 0$, $f^U(\boldsymbol{x}) \geq 0$, $f_2^L(\boldsymbol{\beta}) \leq 0$ and $f_2^U(\boldsymbol{\beta}) \geq 0$,
7. $f_1^L(\boldsymbol{x}) \leq 0$, $f^U(\boldsymbol{x}) \geq 0$ and $f_2^U(\boldsymbol{\beta}) \leq 0$,
8. $f^U(\boldsymbol{x}) \leq 0$, $f^L(\boldsymbol{\beta}) \leq 0$ and $f^U(\boldsymbol{\beta}) \geq 0$,
9. $f_1^U(\boldsymbol{x}) \leq 0$ and $f_2^U(\boldsymbol{\beta}) \leq 0$.

We can assume without loss of generality that each one these cases occur for all $x \in \mathcal{X}$ and $\boldsymbol{\beta} \in \mathcal{C}$ since the general case can be handled by considering the 9 different subsets of $\mathcal{X} \times \mathcal{C}$. In the proof, we will restrict ourselves to making the calculations explicit for cases 1 and 2 since the remaining cases can be handled very similarly. Then, $f_1^L f_2^L \leq f \leq f_1^U f_2^U$. Also, we have that

$$f_1^U f_2^U - f_1^L f_2^L = \Big(f_1^U - f_1^L\Big) f_2^U + f_1^L \Big(f_2^U - f_2^L\Big).$$

Recall that $M = \max(A, B)$. Then, it follows that

$$\begin{aligned} \int_{\mathcal{X}} \Big(f_1^U f_2^U - f_1^L f_2^L\Big)^2 dP_0 &\leq 2M \left(\int_{\mathcal{X}} \Big(f_1^U - f_1^L\Big)^2 dP_0(\boldsymbol{x}) + \|f_2^U - f_2^L\|_\infty^2 \right) \\ &\leq 4M\epsilon^2. \end{aligned}$$

This in turn implies that $H_B(\epsilon, \mathcal{F}, \|\cdot\|_{P_0}) \leq H_B(C\epsilon, \mathcal{F}_1, \|\cdot\|_{P_0}) + H_B(C\epsilon, \mathcal{F}_2, \|\cdot\|_\infty)$ with $C = (2M)^{-1}$. Now we consider case 2. It is not difficult to show that

$$f_2^L f_1^U \leq f \leq f_1^U f_2^U.$$

Hence,

$$\int_{\mathcal{X}} \Big(f_1^U f_2^U - f_2^L f_1^U\Big)^2 dP_0 \leq A^2 \|f_2^U - f_2^L, \|_\infty^2 \leq A^2 \epsilon^2$$

and we can take $C = A^{-1}$. □

Lemma C.4.2. *Let $\mathcal{F}$ be a class of functions satisfying $H_B(\epsilon, \mathcal{F}, \|\cdot\|_{P_0}) < \infty$ for every $\epsilon \in (0, \epsilon_0)$ for some given $\epsilon_0 > 0$. If $\mathcal{D} = \mathcal{F} - \mathcal{F}$ the class of all differences of elements of $\mathcal{F}$, then*

$$H_B(\epsilon, \mathcal{D}, \|\cdot\|_{P_0}) \leq 2H_B(\epsilon/2, \mathcal{F}, \|\cdot\|_{P_0}).$$

Proof. Let $\epsilon \in (0, \epsilon_0)$ and $d = f_2 - f_1$ denote an element in $\mathcal{D}$ with $(f_1, f_2) \in \mathcal{F}^2$. Also, let (f_1^L, f_1^U) and (f_2^L, f_2^U) ϵ-brackets for f_1 and f_2. Define $d^L = f_2^L - f_1^U$ and $d^U = f_2^U - f_1^L$. It is clear that (d^L, d^U) is a bracket for d. Furthermore, we have that

$$\begin{aligned}
&\int_{\mathcal{X}} \left(d^U(\boldsymbol{x}, y) - d^L(\boldsymbol{x}, y)\right)^2 dP_0(\boldsymbol{x}, y) \\
&\leq 2\left\{\int_{\mathcal{X}} \left(f_1^U(\boldsymbol{x}, y) - f_1^L(\boldsymbol{x}, y)\right)^2 dP_0(\boldsymbol{x}, y) + \int_{\mathcal{X}} \left(f_2^U(\boldsymbol{x}, y) - f_2^L(\boldsymbol{x}, y)\right)^2 dP_0(\boldsymbol{x}, y)\right\} \\
&\leq 4\epsilon^2.
\end{aligned}$$

Thus,

$$\exp\Big(H_B(2\epsilon, \mathcal{D}, \|\cdot\|_{P_0})\Big) \leq \exp\Big(H_B(\epsilon, \mathcal{F}, \|\cdot\|_{P_0})\Big)^2,$$

which is equivalent to the statement of the lemma. □

Consider the class $\mathcal{G}_{RK}$ defined as

$$\mathcal{G}_{RK} = \Big\{g : g(\boldsymbol{x}) = g_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \psi(\boldsymbol{\alpha}^T\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}, (\psi, \boldsymbol{\alpha}) \in \mathcal{M}_{RK} \times \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)\Big\}. \quad \text{(C.4.1)}$$

where $\mathcal{M}_{RK}$ is the same class defined in (C.1.2).

Lemma C.4.3. *There exists $A > 0$ such that for $\epsilon \in (0, K)$ we have that*

$$H_B(\epsilon, \mathcal{G}_{RK}, \|\cdot\|_{P_0}) \leq \frac{AK}{\epsilon}.$$

Proof. See the proof of Lemma 4.9 in Balabdaoui et al. (2016). □

Lemma C.4.4. *For some constants $C > 0$ and $\delta > 0$ consider the class of functions*

$$\begin{aligned}
\mathcal{D}_{RC\delta} = \Big\{d : d = f_{1,\boldsymbol{\alpha}} - f_{2,\boldsymbol{\alpha}},\ (f_{1,\boldsymbol{\alpha}}, f_{2,\boldsymbol{\alpha}}) \in \mathcal{G}_{RC}^2, \\
\|d(\alpha^T\cdot)\|_{P_0} \leq \delta \text{ for all } \boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)\Big\}.
\end{aligned}$$

Let $\mathcal{H}_{RCv}$ be a class of functions such that

$$\begin{aligned}
\mathcal{H}_{RCv} = \Big\{h : h(\boldsymbol{x}, y) = yd_1(\boldsymbol{\alpha}^T\boldsymbol{x}) - d_2(\boldsymbol{\alpha}^T\boldsymbol{x}),\ (\boldsymbol{x}, y, \boldsymbol{\alpha}) \in \mathcal{X} \times \mathbb{R} \times \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0), \\
(d_1, d_2) \in \mathcal{D}_{RCv}^2\Big\},
\end{aligned}$$

where $C \geq K_0 \vee 1$. Then, for all $\epsilon \in (0, C)$ we have that

$$H_B\Big(\epsilon, \widetilde{\mathcal{H}}, \|\cdot\|_{B,P_0}\Big) \leq H_B\Big(\epsilon \tilde{C}^{-1}, \mathcal{H}_{RCv}, \|\cdot\|_{P_0}\Big) \leq \frac{\tilde{C}C}{\epsilon} \asymp \frac{1}{\epsilon},$$

and

$$\|\tilde{h}\|_{B,P_0} \lesssim \tilde{D}^{-1}v,$$

where

$$A' = A\Big(2(a_0M_0+1)\Big)^{-1/2}, \quad \tilde{D} = 16M_0C \quad and$$
$$\tilde{C} = \frac{1}{8M_0}\left(2a_0 + \frac{1}{2}e^{(2M_0)^{-1}}\right)^{1/2}\frac{1}{C}, \tag{C.4.2}$$

with a_0, M_0 the same constants from Assumption A6, A the same constant in Lemma C.4.3, and $\widetilde{\mathcal{H}} \stackrel{\text{def}}{=} \mathcal{H}_{RCv}\tilde{D}^{-1}$.

Proof. Consider (d_1^L, d_1^U) and (d_2^L, d_2^U) to be ϵ-brackets of the functions $\boldsymbol{x} \mapsto d_1(\boldsymbol{\alpha}^T\boldsymbol{x})$ and $\boldsymbol{x} \mapsto d_2(\boldsymbol{\alpha}^T\boldsymbol{x})$ and some $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$. It follows from Lemma 4.9 of Balabdaoui et al. (2016) and Lemma C.4.2 that there exists some constant $A > 0$ such that

$$H_B\Big(\epsilon, \mathcal{D}_{RC}, \|\cdot\|_{P_0}\Big) \leq \frac{AC}{\epsilon}.$$

Define now

$$h^L(\boldsymbol{x}, y) = \begin{cases} yd_1^L(\boldsymbol{x}) - d_2^U(\boldsymbol{x}), & \text{if } y \geq 0 \\ yd_1^U(\boldsymbol{x}) - d_2^U(\boldsymbol{x}), & \text{if } y < 0, \end{cases}$$

and

$$h^U(\boldsymbol{x}, y) = \begin{cases} yd_1^U(\boldsymbol{x}) - d_2^L(\boldsymbol{x}), & \text{if } y \geq 0 \\ yd_1^L(\boldsymbol{x}) - d_2^L(\boldsymbol{x}), & \text{if } y < 0. \end{cases}$$

Note first that (h^L, h^U) is a bracket for $h(\boldsymbol{x}, y) = yd_1(\boldsymbol{\alpha}^T\boldsymbol{x}) - d_2(\boldsymbol{\alpha}^T\boldsymbol{x})$. Next we compute the size of this bracket with respect to $\|\cdot\|_{P_0}$. We have that

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathbb{R}} \Big(h^U(\boldsymbol{x}, y) - h^L(\boldsymbol{x}, y)\Big)^2 dP_0(\boldsymbol{x}, y) \\
&\leq 2\left\{\int_{\mathcal{X}\times\mathbb{R}} y^2\Big(d_1^U(\boldsymbol{x}) - d_1^L(\boldsymbol{x})\Big)^2 dP_0(\boldsymbol{x}, y) + \int_{\mathcal{X}}\Big(d_2^U(\boldsymbol{x}) - d_2^L\boldsymbol{x})\Big)^2 dG(\boldsymbol{x})\right\} \\
&= 2\Big\{2a_0 \int_{\mathcal{X}}\Big(d_1^U(\boldsymbol{x}) - d_1^L(\boldsymbol{x})\Big)^2 dG(\boldsymbol{x}) + \int_{\mathcal{X}}\Big(d_2^U(\boldsymbol{x}) - d_2^L\boldsymbol{x})\Big)^2 dG(\boldsymbol{x})\Big\} \\
&\leq 2\big(2a_0 + 1\big)\epsilon^2,
\end{aligned}$$

where a_0 is the same constant of Assumption A6. It follows that

$$H_B\Big(\epsilon, \mathcal{H}, \|\cdot\|_{P_0}\Big) \le \frac{\tilde{A}C}{\epsilon},$$

with $\tilde{A} = A\Big(2(2a_0+1)\Big)^{-1/2}$ and A is the same constant of Lemma C.4.3. Let now $D > 0$ be some constant to be determined later. For a given $h \in \mathcal{H}_{RK^2v}$, we consider $\tilde{h} = D^{-1}h$ which admits $[D^{-1}h^L, D^{-1}h^U]$ as bracket. We will compute the size of this bracket with respect to the Bernstein norm. By definition of the latter we can write for any function h such that h^k is P_0 integrable that

$$\|h\|^2_{B,P_0} = 2\sum_{k=2}^{\infty}\frac{1}{k!}|h|^k dP_0.$$

Thus, using this and convexity of the function $x \mapsto |x|^k$ for all $k \ge 2$ it follows that

$$\begin{aligned}
&\|D^{-1}h^U - D^{-1}h^L\|^2_{B,P_0} \\
&= 2\sum_{k=2}^{\infty}\frac{1}{k!D^k}\int_{\mathcal{X}\times\mathbb{R}}\Big|y\big(d_1^U(\boldsymbol{x}) - d_1^L(\boldsymbol{x})\big) + d_2^U(\boldsymbol{x}) - d_2^L(\boldsymbol{x})\Big|^k dP_0(\boldsymbol{x},y) \\
&\le 2\sum_{k=2}^{\infty}\frac{2^{k-1}}{k!D^k}\bigg\{\int_{\mathcal{X}\times\mathbb{R}}|y|^k\big(d_1^U(\boldsymbol{x}) - d_1^L(\boldsymbol{x})\big)^k dP_0(\boldsymbol{x},y) \\
&\qquad\qquad + \int_{\mathcal{X}\times\mathbb{R}}\big(d_2^U(\boldsymbol{x}) - d_2^L(\boldsymbol{x})\big)^k dP_0(\boldsymbol{x},y)\bigg\}.
\end{aligned}$$

Using Assumption A7 and the fact that $|d_i^L| \le K^2$ and $|d_i^U| \le 2C$ for $i \in \{1,2\}$ (an assumption that one can always make in constructing brackets for a bounded class) we can write

$$\begin{aligned}
&\|D^{-1}h^U - D^{-1}h^L\|^2_{B,P_0} \\
&\le \sum_{k=2}^{\infty}\frac{1}{k!}\left(\frac{2}{D}\right)^k\bigg\{a_0 M_0^{k-2}k!(4C)^{k-2}\int_{\mathcal{X}}\big(d_1^U(\boldsymbol{x}) - d_1^L(\boldsymbol{x})\big)^2 dP_0(\boldsymbol{x},y) \\
&\qquad\qquad + (4C)^{k-2}\int_{\mathcal{X}}\big(d_1^U(\boldsymbol{x}) - d_1^L(\boldsymbol{x})\big)^2 dP_0(\boldsymbol{x},y)\bigg\} \\
&= \left(\frac{2}{D}\right)^2\left\{a_0\sum_{k=2}^{\infty}\left(\frac{8M_0C}{D}\right)^{k-2} + \sum_{k=2}^{\infty}\frac{1}{k!}\left(\frac{8C}{D}\right)^{k-2}\right\}\epsilon^2 \\
&\le \left(\frac{2}{D}\right)^2\left\{a_0\sum_{k=0}^{\infty}\left(\frac{8M_0C}{D}\right)^{k} + \frac{1}{2}\sum_{k=0}^{\infty}\frac{1}{k!}\left(\frac{8C}{D}\right)^{k}\right\}\epsilon^2,
\end{aligned}$$

using $k! \ge 2(k-2)!$. Taking $D = \tilde{D} = 16M_0C$ yields

$$\|\tilde{D}^{-1}h^U - \tilde{D}^{-1}h^L\|^2_{B,P_0} \le \left(\frac{2}{\tilde{D}}\right)^2\left(2a_0 + \frac{1}{2}e^{(2M_0)^{-1}}\right)\epsilon^2,$$

which in turn implies that

$$\|\tilde{D}^{-1}h^U - \tilde{D}^{-1}h^L\|_{B,P_0} \le \frac{1}{8M_0}\left(2a_0 + \frac{1}{2}e^{(2M_0)^{-1}}\right)^{1/2} \frac{1}{C}\ \epsilon.$$

This completes the proof of the first claim about the entropy bound of the class $\widetilde{\mathcal{H}}$ with $\tilde{D}$ defined as above. Now for a given element $\tilde{h} \in \widetilde{\mathcal{H}}$ we calculate

$$\begin{aligned}
\|\tilde{h}\|^2_{B,P_0} &= 2\sum_{k=2}^{\infty} \frac{1}{\tilde{D}^k}\frac{1}{k!}\int_{\mathcal{X}\times\mathbb{R}} \left|yd_1(\boldsymbol{\alpha}^T\boldsymbol{x}) - d_2(\boldsymbol{\alpha}^T\boldsymbol{x})\right|^k dP_0(\boldsymbol{x},y) \\
&\le 2\sum_{k=2}^{\infty} \frac{2^{k-1}}{\tilde{D}^k}\frac{1}{k!}\int_{\mathcal{X}\times\mathbb{R}} \left\{|y|^k\left|d_1(\boldsymbol{\alpha}^T\boldsymbol{x})\right|^k + \left|d_2(\boldsymbol{\alpha}^T\boldsymbol{x})\right|^k dP_0(\boldsymbol{x},y)\right\} \\
&\le 2\sum_{k=2}^{\infty} \frac{2^{k-1}}{\tilde{D}^k}\frac{1}{k!}(2C)^{k-2}\Big\{a_0 M_0^{k-2}k!\int_{\mathcal{X}\times\mathbb{R}} \left|d_1(\boldsymbol{\alpha}^T\boldsymbol{x})\right|^2 dP_0(\boldsymbol{x},y) \\
&\qquad\qquad + \int_{\mathcal{X}\times\mathbb{R}} \left|d_2(\boldsymbol{\alpha}^T\boldsymbol{x})\right|^2 dP_0(\boldsymbol{x},y)\Big\} \\
&\le \left(\frac{2}{\tilde{D}}\right)^2 \left\{a_0\sum_{k=2}^{\infty}\left(\frac{8M_0C}{\tilde{D}}\right)^{k-2} + \sum_{k=2}^{\infty}\frac{1}{k!}\left(\frac{8C}{\tilde{D}}\right)^{k-2}\right\} v^2 \\
&\le \left(\frac{2}{\tilde{D}}\right)^2 \left(2a_0 + \frac{1}{2}e^{(2M_0)^{-1}}\right)\ v^2,
\end{aligned}$$

implying that

$$\|\tilde{h}\|_{B,P_0} \le 2\left(2a_0 + \frac{1}{2}e^{(2M_0)^{-1}}\right)^{1/2}\frac{1}{\tilde{D}}\ v \lesssim \tilde{D}^{-1}v,$$

as claimed. □

Recall that $\mathcal{X}$ is the support of the covariates $X_i, 1 \le i \le n$. Let us denote by $\mathcal{X}_j,\ 1 \le j \le d$ the set of the j-th projection of $\boldsymbol{x} \in \mathcal{X}$. Also, consider some function s that $d-1$ times continuously differentiable on a convex and bounded set $\mathcal{C} \in \mathbb{R}^{d-1}$ with a nonempty interior such that there exists $M > 0$ satisfying

$$\max_{k.\le d-1}\sup_{\boldsymbol{\beta}\in\mathcal{C}} |D^k s(\boldsymbol{\beta})| \le M, \tag{C.4.3}$$

where $k = (k_1, \ldots, k_d)$ with k_j an integer $\in \{0, \ldots, d-1\}$, $k. = \sum_{i=1}^{d-1} k_i$ and

$$D^k \equiv \frac{\partial^{k\cdot} s(\boldsymbol{\beta})}{\partial\beta_{k_1}\ldots\partial\beta_{k_d}}.$$

Consider now the class

$$\begin{aligned}
\mathcal{Q}_{jRC} = \Big\{&q_j(\boldsymbol{x},y) = s(\boldsymbol{\beta})x_j(y - \psi(\boldsymbol{\alpha}^T\boldsymbol{x})), \\
&(\boldsymbol{\alpha},\boldsymbol{\beta},\psi) \in \mathcal{B}(\alpha_0,\delta_0)\times\mathcal{C}\times\mathcal{M}_{RC}, (x_j,y) \in \mathcal{X}_j\times \in \mathbb{R}\Big\}.
\end{aligned} \tag{C.4.4}$$

Define

$$\widetilde{\mathcal{Q}}_{jRC} = \Big\{ \tilde{q}_j : \tilde{q}_j = q_j \tilde{D}^{-1}, q_j \in \mathcal{Q}_{RC} \Big\},$$

where $\tilde{D} > 0$ is some appropriate constant.

Lemma C.4.5. *Let $\epsilon \in (0,1)$ and $C \geq \max(1, 2M_0, Me^{-1/4}2^{-1/2}R^{-1}, 2a_0^{1/2}e^{-1/2})$. Then, there exist some constant $B_1 > 0$ and B_2 depending on a_0, M_0, and R such that*

$$H_B\Big(\epsilon, \widetilde{\mathcal{Q}}_{jRC}, \|\cdot\|_{B,P_0}\Big) \leq \frac{B_1 C}{\epsilon}, \quad \|\tilde{q}_j\|_{B,P_0} \leq B_2,$$

if $\tilde{D} = 8MRC$ where a_0 and M_0 are the same positive constants in Assumption A6, and M is from (C.4.3).

Proof. Fix $j \in \{1, \ldots, d\}$. The proof of this lemma uses similar techniques as in showing Lemma C.4.4. Let (g^L, g^U) be ϵ-brackets for the class $\mathcal{G}_{RC}$. Using the result of Lemma C.4.3 we know that there are at most $N \leq \exp(AC/\epsilon)$ such brackets covering $\mathcal{G}_{RC}$ for some constant $A > 0$. Define

$$\Big(k^L(\boldsymbol{x}, y), k^U(\boldsymbol{x}, y)\Big) = \begin{cases} \Big(x_j(y - g^L(\boldsymbol{x})), x_j(y - g^U(\boldsymbol{x}))\Big), & \text{if } x_j \geq 0 \\ \Big(x_j(y - g^U(\boldsymbol{x})), x_j(y - g^L(\boldsymbol{x}))\Big), & \text{if } x_j < 0\ . \end{cases} \tag{C.4.5}$$

Then, the collection of all possible pairs (q^L, q^U) form brackets for the class of functions

$$\mathcal{K}_{jRC} = \Big\{ k_j(\boldsymbol{x}, y) = x_j(y - \psi(\boldsymbol{\alpha}^T \boldsymbol{x})), \\ (\boldsymbol{\alpha}, \psi) \in \mathcal{B}(\alpha_0, \delta_0) \times \mathcal{M}_{RC}, (x_j, \boldsymbol{x}, y) \in \mathcal{X}_j \times \mathcal{X} \times \mathbb{R} \Big\}.$$

Furthermore we have that

$$\begin{aligned} \|k^U - k^L\|_{P_0}^2 &= \int_{\mathcal{X}} x_j^2 \big(g^U(\boldsymbol{x}) - g^L(\boldsymbol{x})\big)^2 dG(\boldsymbol{x}) \\ &\leq \|\boldsymbol{x}\|_2^2 \int_{\mathcal{X}} \big(g^U(\boldsymbol{x}) - g^L(\boldsymbol{x})\big)^2 dG(\boldsymbol{x}) \leq R^2 \epsilon^2. \end{aligned}$$

This implies that

$$H_B\Big(\epsilon, \mathcal{K}_{jRC}, \|\cdot\|_{P_0}\Big) \leq \frac{ARC}{\epsilon},$$

where A is the same constant of Lemma C.4.3. Furthermore, the assumption in (C.4.3) implies that the function s belongs to $C_{\tilde{M}}^{d-1}$ as defined in Section 2.7 in van der Vaart and Wellner (1996), with $\tilde{M} = 2M$. Using now Theorem 2.7.1 of van der Vaart and Wellner (1996) it follows that there exists some constant $B > 0$ such that

$$\log N\Big(\epsilon, C_{\tilde{M}}^{d-1}, \|\cdot\|_\infty\Big) \leq B \left(\frac{1}{\epsilon}\right)^{d/(d-1)} \leq \frac{B}{\epsilon}.$$

This also implies that

$$H_B\Big(\epsilon, C^{d-1}_{\tilde{M}}, \|\cdot\|_\infty\Big) = \log N\Big(\epsilon/2, C^{d-1}_{\tilde{M}}, \|\cdot\|_\infty\Big) \leq \frac{2B}{\epsilon}.$$

Indeed, for an arbitrary $s \in C^{d-1}_{\tilde{M}}$ there exists $s_i, i \in \{1,\dots,N\}$, with $N = N\big(\epsilon/2, C^{d-1}_{\tilde{M}}, \|\cdot\|_\infty\big)$, such that $\|s - s_i\|_\infty \leq \epsilon/2$. The claim follows from noting that $(s_i - \epsilon/2, s_i + \epsilon/2)$ is an ϵ-bracket for $C^{d-1}_{\tilde{M}}$ with respect to $\|\cdot\|_\infty$. Using Lemma C.4.1 it follows that there exists some constant $L > 0$ such that

$$H_B\Big(\epsilon, \mathcal{Q}_{jRC}, \|\cdot\|_{P_0}\Big) \quad \leq \quad L\left(\frac{1}{\epsilon} + \frac{C}{\epsilon}\right) \leq \frac{2LC}{\epsilon}, \tag{C.4.6}$$

using that $C \geq 1$, $d-1 \geq 1$ and $\epsilon \in (0,1)$. Consider now a constant $D > 0$, and (q^L, q^U) and ϵ-bracket. From the proof of Lemma C.4.1 we know that we can restrict attention to the case for example to case 1 assumed to occur for all $(\boldsymbol{x}, \boldsymbol{\beta}) \in \mathcal{X} \times \mathcal{C}$. In such that we have $q^L = s^L k^L$ and $q^U = s^U k^U$ where (s^L, s^U) is an ϵ-bracket for $C^1_{\tilde{M}}$equipped with $\|\cdot\|_\infty$, where the expression of (k^L, k^U) is given in (C.4.5). We can now write

$$\begin{aligned}
&\|D^{-1}q^U - D^{-1}q^L\|^2_{B,P_0} \\
&\quad = 2\sum_{k=2}^{\infty} \frac{1}{k!}\frac{1}{D^k}\int_{\mathcal{X}\times\mathbb{R}} \big|s^U k^U - s^L k^L\big|^k dP_0 \\
&\quad \leq \sum_{k=2}^{\infty} \frac{2^k}{k!}\frac{1}{D^k}\int_{\mathcal{X}\times\mathbb{R}} \Big\{\big|s^U\big(k^U - k^L\big)\big|^k + \big|k^L\big(s^U - s^L\big)\big|^k\Big\} dP_0,
\end{aligned}$$

with

$$\begin{aligned}
\int_{\mathcal{X}\times\mathbb{R}} \big|s^U\big(k^U - k^L\big)\big|^k dP_0 &\leq M^k(2RC)^{k-2}\int_{\mathcal{X}\times\mathbb{R}} \big(k^U - k^L\big)^2 dP_0 \\
&= M^2(2MCR)^{k-2}\epsilon^2,
\end{aligned}$$

where we used the fact that $|s| \leq M$ by assumption of the lemma (implying that we can constructs brackets (s^L, s^U) satisfying the same property), and $k^U - k^U = x_j(g^U - g^L) \leq 2RC$. Also, if we assume without loss of generality that $x_j \geq 0$ is satisfied for all $\boldsymbol{x} \in \mathcal{X}$ we have that

$$\begin{aligned}
&\int_{\mathcal{X}\times\mathbb{R}} \big|k^L\big(s^U - s^L\big)\big|^k dP_0 \\
&\quad \leq (2M)^{k-2}\int_{\mathcal{X}\times\mathbb{R}} \big|x_j(y - g^L(\boldsymbol{x}))\big|^k dP_0(\boldsymbol{x}, y) \times \epsilon^2 \\
&\quad \leq (2M)^{k-2}R^k 2^{k-1}\int_{\mathcal{X}\times\mathbb{R}} \Big\{|y|^k + \big|g^L(\boldsymbol{x})\big|^k\Big\} dP_0(\boldsymbol{x}, y) \times \epsilon^2. \\
&\quad \leq (2M)^{k-2}R^k 2^{k-1}\Big(a_0 M_0^{k-2} k! + C^k\Big)\epsilon^2.
\end{aligned}$$

Putting these inequalities together and after some algebra we get

$$\|D^{-1}q^U - D^{-1}q^L\|^2_{B,P_0} \leq \Bigg(\frac{1}{2}\left(\frac{2M}{D}\right)^2 e^{4MCR/D} + \left(\frac{2RC}{D}\right)^2 e^{8MCR/D} \\ + 2a_0\left(\frac{2R}{D}\right)^2 \frac{1}{1-8MM_0R/D})\Bigg)\epsilon^2.$$

Now let us choose $\tilde{D} = D \geq \max(16MM_0R, 8MRC)$. In particular, we can assume that C is large enough so that $\max(16MM_0R, 8MRC) = 8MRC = \tilde{D}$ (or equivalently $C \geq 2M_0$). Then, $4MCR/\tilde{D} = 1/2$, $8MCR/\tilde{D} = 1/4$, and $8MM_0R/\tilde{D} = M_0/C \leq 1/2$. Therefore,

$$\begin{aligned}\|\tilde{D}^{-1}q^U - \tilde{D}^{-1}q^L\|^2_{B,P_0} &\leq \left(\frac{1}{2}\left(\frac{2M}{\tilde{D}}\right)^2 e^{1/2} + \left(\frac{2RC}{\tilde{D}}\right)^2 e + 4a_0\left(\frac{2R}{\tilde{D}}\right)^2\right)\ \epsilon^2 \\ &= \Big(2M^2e^{1/2} + 4R^2e\ C^2 + 16a_0R^2\Big)\ \frac{1}{\tilde{D}^2}\ \epsilon^2 \\ &\leq \frac{\tilde{A}C^2}{\tilde{D}^2}\ \epsilon^2 = \frac{\tilde{A}}{64M^2R^2}\epsilon^2,\end{aligned}$$

if C is large enough, where $\tilde{A} = 2M^2e^{1/2} + 4R^2e + 16a_0R^2$. It follows that we can find some constant $\tilde{L} > 0$ such that

$$\|\tilde{D}^{-1}q^U - \tilde{D}^{-1}q^L\|_{B,P_0} \leq \tilde{L}\epsilon.$$

This in turn implies that

$$H_B\big(\tilde{L}\epsilon, \widetilde{Q}_{jRC}, \|\cdot\|_{B,P_0}\big) \leq H_B\big(\epsilon, Q_{jRC}, \|\cdot\|_{P_0}\big) \lesssim \frac{2MC}{\epsilon},$$

using (C.4.6). Hence, we can find a constant $B_1 > 0$ such that

$$H_B\big(\epsilon, \widetilde{Q}_{jRC}, \|\cdot\|_{B,P_0}\big) \leq \frac{B_1C}{\epsilon}.$$

Now we turn to computing an upper bound for $\|\tilde{q}_j\|_{B,P_0}$. We have

$$\begin{aligned}\|\tilde{q}_j\|^2_{B,P_0} &= 2\sum_{k=2}^{\infty}\frac{1}{k!}D^{-k}\int_{\mathcal{X}\times\mathbb{R}} |s(\boldsymbol{\beta})|^k \big|x_j\big(y - \psi(\boldsymbol{\alpha}^T\boldsymbol{x})\big)\big|^k dP_0(\boldsymbol{x}, y) \\ &\leq \sum_{k=2}^{\infty}\frac{1}{k!}2^kD^{-k}(RM)^k\int_{\mathcal{X}\times\mathbb{R}}\Big\{\big|y\big|^k + \big|\psi(\boldsymbol{\alpha}^T\boldsymbol{x})\big|^k\Big\}dP_0(\boldsymbol{x}, y) \\ &\leq \sum_{k=2}^{\infty}\frac{1}{k!}2^kD^{-k}(RM)^k\Big(a_0M_0^{k-2}k! + C^k\Big)\end{aligned}$$

$$\leq a_0 \left(\frac{2MR}{D}\right)^2 \sum_{k=2}^{\infty} \left(\frac{2RMM_0}{D}\right)^{k-2} + \frac{1}{2}\left(\frac{2MRC}{D}\right)^2 \sum_{k=2}^{\infty} \frac{1}{(k-2)!}\left(\frac{2RMC}{D}\right)^{k-2}$$
$$\leq a_0 \left(\frac{1}{2C}\right)^2 \sum_{k=0}^{\infty} \left(\frac{1}{4}\right)^k + \frac{1}{2}\left(\frac{1}{2}\right)^2 \sum_{k=0}^{\infty} \frac{1}{k!}\left(\frac{1}{2}\right)^k$$
$$\leq a_0 \left(\frac{1}{2C}\right)^2 \frac{3}{4} + \frac{1}{2}\left(\frac{1}{2}\right)^2 e^{1/2},$$

if $D = 4MRC$ and $C \geq \max(1, 2M_0)$. The proof of the lemma is complete if we write $B_2 = (3a_0/16 + e^{1/2}/8)^{1/2}$.

□

In the next lemma, we consider a given a class of functions $\mathcal{F}$ which admits a bounded bracketing entropy with respect to $\|\cdot\|_{P_0}$ for $\epsilon \in (0,1]$. Suppose also that there exists $D > 0$ such that $\|f\|_\infty \leq D$ and $\delta > 0$ such that $\|f\|_{P_0} \leq \delta$ for all $f \in \mathcal{F}$. Then we can derive an upper bound for the bracketing entropy for the class

$$\widetilde{\mathcal{F}} = \Big\{\tilde{f} : \tilde{f}(\boldsymbol{x}, y) = (4M_0 D)^{-1} f(\boldsymbol{x})\Big(y - \lambda\psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{x})\Big), (\boldsymbol{x}, y) \in \mathcal{X} \times \mathbb{R} \text{ and } f \in \mathcal{F}\Big\}, \tag{C.4.7}$$

with respect to the Bernstein norm. Here, M_0 is the same constant from Assumption A6 and $\tilde{D}$ is a positive constant that will be determined below.

Lemma C.4.6. *Let $\mathcal{F}$ be a class of functions satisfying the conditions above. Then,*

$$H_B(\epsilon, \widetilde{\mathcal{F}}, \|\cdot\|_{B,P_0}) \leq H_B(\epsilon \tilde{D}^{-1}, \mathcal{F}, \|\cdot\|_{P_0}), \quad \textit{and} \quad \|\tilde{f}\|_{B,P_0} \leq \tilde{D}\delta,$$

where

$$\tilde{D} = \left(\frac{a_0}{2M_0^2} + \frac{\lambda^2 K_0^2}{8M_0^2} e^{\lambda K_0 (2M_0)^{-1}}\right)^{1/2} D^{-1}, \tag{C.4.8}$$

and a_0, M_0 are the same constants from Assumption A6.

Proof. Let (L, U) be an ϵ-bracket for $\mathcal{F}$ with respect to $\|\cdot\|_{P_0}$. Consider the class

$$\mathcal{F}' = \Big\{f' : f'(\boldsymbol{x}, y) = f(\boldsymbol{x})\Big(y - \lambda\psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{x})\Big), (\boldsymbol{x}, y) \in \mathcal{X} \times \mathbb{R} \text{ and } f \in \mathcal{F}\Big\}.$$

Then for $f' \in \mathcal{F}'$ we have if $y - \lambda\psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{x}) \geq 0$

$$L(\boldsymbol{x})(y - \lambda\psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{x})) \leq f'(\boldsymbol{x}, y) \leq U(\boldsymbol{x})(y - \lambda\psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{x})),$$

or, if $y - \lambda\psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{x}) < 0$,

$$U(\boldsymbol{x})(y - \lambda\psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{x})) \leq f'(\boldsymbol{x}, y) \leq L(\boldsymbol{x})(y - \lambda\psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{x})).$$

Let (L', U') denote the new bracket. Using the definition of the Bernstein norm, convexity of $x \mapsto x^k,\ k \geq 2$ and $\|\psi_0\|_\infty \leq K_0$ we have that

$$\begin{aligned}
&\left\|(U' - L')(4M_0D)^{-1}\right\|^2_{B,P_0} \\
&= 2\sum_{k=2}^{\infty} \frac{(4M_0D)^{-k}}{k!} \int_{\mathcal{X}\times\mathbb{R}} (U(\boldsymbol{x}) - L(\boldsymbol{x}))^k |y - \lambda\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{x})|^k dP_0(\boldsymbol{x}, y) \\
&\leq 2\sum_{k=2}^{\infty} \frac{(4M_0D)^{-k}}{k!} \int_{\mathcal{X}\times\mathbb{R}} (U(\boldsymbol{x}) - L(\boldsymbol{x}))^k 2^{k-1}\Big(|y|^k + \lambda^k|\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{x})|^k\Big) dP_0(\boldsymbol{x}, y) \\
&\leq \sum_{k=2}^{\infty} \frac{1}{2^kD^kM_0^kk!} \int_{\mathcal{X}\times\mathbb{R}} (U(\boldsymbol{x}) - L(\boldsymbol{x}))^k \big(|y|^k + \lambda^kK_0^k\big) dP_0(\boldsymbol{x}, y) \\
&\leq \sum_{k=2}^{\infty} \frac{1}{2^kD^kM_0^kk!} \int_{\mathcal{X}\times\mathbb{R}} (U(\boldsymbol{x}) - L(\boldsymbol{x}))^k (a_0k!M_0^{k-2} + \lambda^kK_0^k) dP_0(\boldsymbol{x}, y) \\
&\leq \frac{a_0}{4M_0^2D^2}\sum_{k=2}^{\infty} \frac{1}{2^{k-2}} \int_{\mathcal{X}\times\mathbb{R}} (U(\boldsymbol{x}) - L(\boldsymbol{x}))^2 g(\boldsymbol{x})dx \\
&\qquad\qquad + \frac{\lambda^2K_0^2}{4D^2M_0^2}\sum_{k=2}^{\infty}\left(\frac{\lambda K_0}{2M_0}\right)^{k-2}\frac{1}{k!}\int_{\mathcal{X}\times\mathbb{R}} (U(\boldsymbol{x}) - L(\boldsymbol{x}))^2 g(\boldsymbol{x})d\boldsymbol{x} \\
&\leq \frac{a_0}{2M_0^2D^2}\epsilon^2 + \frac{\lambda^2K_0^2}{8D^2M_0^2}\sum_{k=2}^{\infty}\frac{1}{(k-2)!}\left(\frac{\lambda K_0}{2M_0}\right)^{k-2}\ \epsilon^2 \\
&\leq \left(\frac{a_0}{2M_0^2D^2} + \frac{\lambda^2K_0^2}{8D^2M_0^2}e^{\lambda K_0(2M_0)^{-1}}\right)\epsilon^2 = \tilde{D}^2\ \epsilon^2.
\end{aligned}$$

This implies that

$$H_B\Big(\epsilon\tilde{D}, \widetilde{\mathcal{F}}, \|\cdot\|_{B,P_0}\Big) \leq H_B\Big(\epsilon, \mathcal{F}, \|\cdot\|_{P_0}\Big),$$

or equivalently

$$H_B\Big(\epsilon, \mathcal{F}', \|\cdot\|_{B,P_0}\Big) \leq H_B\Big(\epsilon\tilde{D}^{-1}, \mathcal{F}, \|\cdot\|_{P_0}\Big).$$

Using similar calculations we can write

$$\begin{aligned}
&\|\tilde{f}\|^2_{B,\mathbb{P}} = 2\sum_{k=2}^{\infty}\frac{1}{(4M_0D)^k}\frac{1}{k!}\int_{\mathcal{X}\times\mathbb{R}} |f(\boldsymbol{x})|^k|y - \lambda\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{x})|^k dP_(\boldsymbol{x}, y) \\
&\leq \frac{1}{D^2}\sum_{k=2}^{\infty}\frac{1}{(2M_0)^k}\frac{1}{k!}\int_{\mathcal{X}\times\mathbb{R}} f(\boldsymbol{x})^2(a_0k!M_0^{k-2} + \lambda^kK_0^k)g(\boldsymbol{x})d\boldsymbol{x} \\
&\leq \left(\frac{a_0}{4M_0^2D^2}\sum_{k=2}^{\infty}\frac{1}{2^{k-2}} + \frac{\lambda^2K_0^2}{8D^2M_0^2}\sum_{k=2}^{\infty}\left(\frac{\lambda K_0}{2M_0}\right)^{k-2}\frac{1}{(k-2)!}\right)\int_{\mathcal{X}\times\mathbb{R}} f(\boldsymbol{x})^2g(\boldsymbol{x})d\boldsymbol{x} \\
&\leq \tilde{D}^2\delta^2,
\end{aligned}$$

which completes the proof. □

In the next corollary, we consider the class

$$\mathcal{F} = \Big\{ x \mapsto f_{\boldsymbol{\alpha}}(\boldsymbol{x}) = E_{i,\boldsymbol{\alpha}_0}(\boldsymbol{\alpha}_0^T \boldsymbol{x}) - E_{i,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{x}),\ \boldsymbol{x} \in \mathcal{X}, \boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta) \Big\},$$

where $E_{i,\boldsymbol{\alpha}}(u) = \mathbb{E}\left\{X_i | \boldsymbol{\alpha}^T \boldsymbol{X} = u\right\}$ for $i \in \{1, \ldots, d\}$ and $\delta \in (0, \delta_0)$. Using the same arguments in the proof of Lemma C.5.3 in Section C.5 with $f(\boldsymbol{x}) = x_i$ it follows that for all $x \in \mathcal{X}$ and $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta)$

$$|f_{\boldsymbol{\alpha}'}(\boldsymbol{x}) - f_{\boldsymbol{\alpha}}(\boldsymbol{x})| \le M \|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\|,$$

for the same constant M of that lemma. Now, we can apply Theorem 2.7.11 of van der Vaart and Wellner (1996) to conclude that

$$N_B\big(2\epsilon, \mathcal{F}, \|\cdot\|_{P_0}\big) \le N\big(\epsilon, \mathcal{B}(\boldsymbol{\alpha}_0, \delta), \|\cdot\|\big),$$

where $N\big(\epsilon, \mathcal{B}(\boldsymbol{\alpha}_0, \delta), \|\cdot\|\big)$ is the ϵ-covering number for $\mathcal{B}(\boldsymbol{\alpha}_0, \delta)$ with respect to the norm $\|\cdot\|$ which is of order $(\delta/\epsilon)^d$ for $\epsilon \in (0, \delta)$. Hence, using the inequality $\log(\boldsymbol{x}) \le x$ for $x > 0$ we can find a constant $M' > 0$ depending on d such that

$$H_B\big(\epsilon, \mathcal{F}, \|\cdot\|_{P_0}\big) \le \frac{M'\delta}{\epsilon}.$$

Furthermore, there exists $\tilde{M} > 0$ such that $\|f\|_\infty \le \tilde{M}\delta$ and $\|f\|_{P_0} \le \tilde{M}\delta$.

Lemma C.4.7. *Let $\mathcal{F}$ be the class of functions as above and consider the related class*

$$\mathcal{F}' = \Big\{ f' : f'(\boldsymbol{x}, y) = f(\boldsymbol{x})\Big(y - \lambda\psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{x})\Big), (\boldsymbol{x}, y) \in \mathcal{X} \times \mathbb{R}, f \in \mathcal{F} \Big\}. \qquad \text{(C.4.9)}$$

Then,

$$E[\|\mathbb{G}_n\|_{\mathcal{F}'}] \lesssim \delta.$$

Proof. Note that for any function $f' \in \mathcal{F}'$ and constant $C > 0$ we have that $\mathbb{G}_n(f'C^{-1}) = C^{-1}\mathbb{G}_n f'$ implying that $\|\mathbb{G}_n\|_{\mathcal{F}'} = 4M_0\tilde{M}\delta\|\mathbb{G}_n\|_{\widetilde{\mathcal{F}}}$, where

$$\widetilde{\mathcal{F}} = \Big\{ \tilde{f} : \tilde{f}(\boldsymbol{x}, y) = (4M_0\tilde{M}\delta)^{-1} f'(\boldsymbol{x}, y), f' \in \mathcal{F}' \Big\}.$$

Note also that the constant $\tilde{D}$ in Lemma C.4.6 is given by $\tilde{D} \asymp \delta^{-1}$, where $\tilde{D}$ depends on $\tilde{M}$, a_0, M_0 and K_0. Also, using the entropy calculations along with Lemma C.4.6 we can show easily that

$$H_B(\epsilon, \widetilde{\mathcal{F}}, \|\cdot\|_{B,P_0}) \lesssim \frac{1}{\epsilon},$$

and that $\|\tilde{f}\|_{B,P_0} \lesssim 1$. Using Lemma 3.4.3 of van der Vaart and Wellner (1996) it follows that there exists some constant $B > 0$ such that

$$E[\|\mathbb{G}_n\|_{\widetilde{\mathcal{F}}}] \lesssim J_n\left(1 + \frac{J_n}{\sqrt{n}B^2}\right),$$

with $J_n = \int_0^B \sqrt{1 + B/\epsilon}\, d\epsilon$. Hence, $E[\|\mathbb{G}_n\|_{\widetilde{\mathcal{F}}}] \lesssim 1$ and $E[\|\mathbb{G}_n\|_{\mathcal{F}'}] \lesssim \delta$ as claimed. □

C.5 Auxiliary results

Proof of Lemma 4.3.1. We have

$$\begin{aligned}&(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\\ &\quad= (\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0).\end{aligned}$$

In the parameterizations that we consider, the columns of $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)$ are orthogonal to $\boldsymbol{\alpha}_0$. We can therefore extend the matrix $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)$ with a last column $\boldsymbol{\alpha}_0$ to a square nonsingular matrix $\bar{\boldsymbol{J}}_{\mathbb{S}}(\boldsymbol{\beta}_0)$. This leads to the equality

$$\begin{aligned}&\left(\bar{\boldsymbol{J}}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\bar{\boldsymbol{J}}_{\mathbb{S}}(\boldsymbol{\beta}_0)\\ &\quad= \left(\bar{\boldsymbol{J}}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T \boldsymbol{A}\bar{\boldsymbol{J}}_{\mathbb{S}}(\boldsymbol{\beta}_0).\end{aligned}$$

Multiplying on the left by $\left(\left(\bar{\boldsymbol{J}}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right)^T\right)^{-1}$ and on the right by $\bar{\boldsymbol{J}}_{\mathbb{S}}(\boldsymbol{\beta}_0)^{-1}$, we get:

$$\boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A} = \boldsymbol{A}. \tag{C.5.1}$$

This shows that $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T$ is a generalized inverse of $\boldsymbol{A}$.

To complete the proof and show that it is indeed the Moore-Penrose inverse of $\boldsymbol{A}$, we first note that

$$\begin{aligned}&\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\\ &\quad= \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T. \end{aligned}\tag{C.5.2}$$

Furthermore,

$$\begin{aligned}&\left(\boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\right)^T\\ &\quad= \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}^T\\ &\quad= \boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A},\end{aligned}$$

where the last equality holds since $\boldsymbol{A}$ is symmetric, being a covariance matrix. We have to show that

$$\begin{aligned}&\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\\ &\quad= \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T \boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T. \end{aligned}\tag{C.5.3}$$

Multiplying on the left by $(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T$ and on the right by $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)$, we get

$$\begin{aligned}
&(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\\
&=(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\\
&=(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0),
\end{aligned}$$

and (C.5.3) follows by the orthogonality relation of the columns of $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)$ with $\boldsymbol{\alpha}_0$ in the same way as before, replacing the matrix $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)$ by $\bar{\boldsymbol{J}}_{\mathbb{S}}(\boldsymbol{\beta}_0)$ in the outer factors of the equality relation. In a similar way we obtain

$$\begin{aligned}
&\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\boldsymbol{A}\right)^T\\
&=\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\boldsymbol{A}. \qquad \text{(C.5.4)}
\end{aligned}$$

Since the matrix $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\boldsymbol{A}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T$ satisfies properties (C.5.1), (C.5.2),(C.5.3) and (C.5.4), the matrix satisfies the four properties which define the Moore-Penrose pseudo-inverse matrix of $\boldsymbol{A}$. This completes the proof of Lemma 4.3.1. □

Remark C.5.1. *The same proof holds for showing that the Moore-Penrose inverse $\tilde{\boldsymbol{A}}$ is given by*

$$\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\left\{(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T\tilde{\boldsymbol{A}}\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)\right\}^{-1}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0))^T.$$

Lemma C.5.1 (Derivative $\boldsymbol{\alpha}\mapsto\psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})$).

$$\frac{\partial}{\partial\alpha_j}\psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0}=\left(x_j-E(X_j|\boldsymbol{\alpha}^T\boldsymbol{X}=\boldsymbol{\alpha}_0^T\boldsymbol{x})\right)\psi_0'(\boldsymbol{\alpha}_0^T\boldsymbol{x}),$$

and

$$\begin{aligned}
\frac{\partial}{\partial\beta_j}\psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0}&=\frac{\partial}{\partial\beta_j}\psi_{\mathbb{S}(\boldsymbol{\beta})}(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\\
&=\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T\right)_j\left(\boldsymbol{x}-E(\boldsymbol{X}|\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{X}=\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x})\right)\psi_0'\left(\mathbb{S}(\boldsymbol{\beta})^T\boldsymbol{x}\right),
\end{aligned}$$

where $\left(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T\right)_j$ denotes the jth row of $\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}_0)^T$.

Proof. We assume without loss of generality that the first component α_1 of $\boldsymbol{\alpha}$ is not equal to zero. Denote the conditional density of $(X_2,\dots,X_d)^T$ given $\boldsymbol{\alpha}^T\boldsymbol{X}=u$ by $h_{\boldsymbol{\alpha}}(\cdot|u)$. Using the change of variables $t_1=\boldsymbol{\alpha}^T\boldsymbol{x}$, $t_j=x_j$ for $1\le j\le d$, the function $\psi_{\boldsymbol{\alpha}}$ can be written as

$$\begin{aligned}\psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x}) &= \mathbb{E}[\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = \boldsymbol{\alpha}^T\boldsymbol{x}] \\ &= \int \psi_0\left(\frac{\alpha_{01}}{\alpha_1}(\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2\tilde{x}_2 - \ldots - \alpha_d\tilde{x}_d) + \sum_{j=2}^{d}\alpha_{0j}\tilde{x}_j\right) \\ &\qquad\qquad \cdot h_{\boldsymbol{\alpha}}(\tilde{x}_2,\ldots,\tilde{x}_d|\boldsymbol{\alpha}^T\boldsymbol{x})\prod_{j=2}^{d}d\tilde{x}_j,\end{aligned}$$

with partial derivatives w.r.t. α_j for $2 \le j \le d$ given by

$$\begin{aligned}&\frac{\partial}{\partial\alpha_j}\psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x}) \\ &= \frac{\partial}{\partial\alpha_j}\mathbb{E}[\psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = \boldsymbol{\alpha}^T\boldsymbol{x}] \\ &= \int \frac{\alpha_{01}}{\alpha_1}(x_j - \tilde{x}_j)\psi_0'\left(\frac{\alpha_{01}}{\alpha_1}(\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2\tilde{x}_2 - \ldots - \alpha_d\tilde{x}_d) + \sum_{j=2}^{d}\alpha_{0j}\tilde{x}_j\right) \\ &\qquad\qquad \cdot h_{\boldsymbol{\alpha}}(\tilde{x}_2,\ldots,\tilde{x}_d|\boldsymbol{\alpha}^T\boldsymbol{x})\prod_{j=2}^{d}d\tilde{x}_j \\ &\quad + \int \psi_0\left(\frac{\alpha_{01}}{\alpha_1}(\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2\tilde{x}_2 - \ldots - \alpha_d\tilde{x}_d) + \sum_{j=2}^{d}\alpha_{0j}\tilde{x}_j\right)\frac{\partial}{\partial\alpha_j} \\ &\qquad\qquad \cdot h_{\boldsymbol{\alpha}}(\tilde{x}_2,\ldots,\tilde{x}_d|\boldsymbol{\alpha}^T\boldsymbol{x})\prod_{j=2}^{d}d\tilde{x}_j,\end{aligned}$$

which is at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ equal to

$$\begin{aligned}\frac{\partial}{\partial a_j}\psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x})\Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0} &= \int (x_j - \tilde{x}_j)\psi_0^T\left(\boldsymbol{\alpha}_0^T\boldsymbol{x}\right)h_{\boldsymbol{\alpha}_0}(\tilde{x}_2,\ldots,\tilde{x}_d|\boldsymbol{\alpha}_0^T\boldsymbol{x})\prod_{j=2}^{d}d\tilde{x}_j \\ &= \psi_0'(\boldsymbol{\alpha}_0^T\boldsymbol{x})\left\{x_j - \mathbb{E}(X_j|\boldsymbol{\alpha}_0^T\boldsymbol{X} = \boldsymbol{\alpha}_0^T\boldsymbol{x})\right\}.\end{aligned}$$

For the partial derivatives w.r.t. α_1 we have

$$\begin{aligned}&\frac{\partial}{\partial\alpha_1}\psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{x}) \\ &= \int\left\{\frac{\alpha_{01}}{\alpha_1}x_1 - \frac{\alpha_{01}}{\alpha_1^2}(\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2\tilde{x}_2 - \ldots - \alpha_d\tilde{x}_d)\right\} \\ &\quad \cdot \psi_0'\left(\frac{\alpha_{01}}{\alpha_1}(\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2\tilde{x}_2 - \ldots - \alpha_d\tilde{x}_d) + \sum_{j=2}^{d}\alpha_{0j}\tilde{x}_j\right)h_{\boldsymbol{\alpha}}(\tilde{x}_2,\ldots,\tilde{x}_d|\boldsymbol{\alpha}^T\boldsymbol{x})\prod_{j=2}^{d}d\tilde{x}_j\end{aligned}$$

$$+\int \psi_0 \left(\boldsymbol{\alpha}^T \boldsymbol{x} + (\alpha_{01} - \alpha_1) \frac{\boldsymbol{\alpha}^T \boldsymbol{x} - \alpha_2 \tilde{x}_2 - \ldots - \alpha_d \tilde{x}_d}{\alpha_1} + \sum_{j=2}^{d} (\alpha_{0j} - \alpha_j) \tilde{x}_j \right)$$
$$\cdot \frac{\partial}{\partial \alpha_1} h(\tilde{x}_2, \ldots, \tilde{x}_d | \boldsymbol{\alpha}^T \boldsymbol{x}) \prod_{j=2}^{d} d\tilde{x}_j,$$

and,

$$\frac{\partial}{\partial a_1} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{x}) \Big|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0} = \psi_0'(\boldsymbol{\alpha}_0^T \boldsymbol{x}) \left\{ x_1 - E(X_1 | \boldsymbol{\alpha}_0^T \boldsymbol{X} = \boldsymbol{\alpha}_0^T \boldsymbol{x}) \right\}.$$

This proves the first result of Lemma C.5.1. The proof for the second results follows similarly and is omitted. □

Lemma C.5.2. *Let $\bar{\xi}$ be defined by*

$$\bar{\xi}(\boldsymbol{\alpha}) = \int \boldsymbol{x} \left\{ y - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{x}) \right\} dP_0(\boldsymbol{x}, y) = \int \boldsymbol{x} \left\{ \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{x}) \right\} dG(\boldsymbol{x}),$$

then we have for each $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$,

$$\bar{\xi}(\boldsymbol{\alpha}) = \mathbb{E}\left[\mathit{Cov}\left[\boldsymbol{X}, \psi_0(\boldsymbol{\alpha}^T \boldsymbol{X} + (\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \boldsymbol{X}) | \boldsymbol{\alpha}^T \boldsymbol{X} \right] \right].$$

Moreover,

$$\boldsymbol{\alpha}^T \bar{\xi}(\boldsymbol{\alpha}) = 0,$$

and,

$$\begin{aligned}(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \bar{\xi}(\boldsymbol{\alpha}) &= \mathbb{E}\left[\mathit{Cov}\left[(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \boldsymbol{X}, \psi_0(\boldsymbol{\alpha}^T \boldsymbol{X} + (\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \boldsymbol{X}) | \boldsymbol{\alpha}^T \boldsymbol{X} \right] \right] \\ &\geq 0,\end{aligned}$$

and $\boldsymbol{\alpha}_0$ is the only value such that the above equation holds uniform in $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$.

Proof. We have,

$$\begin{aligned}\bar{\xi}(\boldsymbol{\alpha}) &= \int \boldsymbol{x} \left\{ y - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{x}) \right\} dP_0(\boldsymbol{x}, y) = \int \boldsymbol{x} \left\{ \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \boldsymbol{x}) \right\} dG(\boldsymbol{x}) \\ &= \int \boldsymbol{x} \left[\psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{x}) - \mathbb{E}\left\{ \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X}) | \boldsymbol{\alpha}^T \boldsymbol{X} = \boldsymbol{\alpha}^T \boldsymbol{x} \right\} \right] dG(\boldsymbol{x}) \\ &= \mathbb{E}\left[\mathrm{Cov}\left[\boldsymbol{X}, \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X}) | \boldsymbol{\alpha}^T \boldsymbol{X} \right] \right],\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\alpha}^T \bar{\xi}(\boldsymbol{\alpha}) &= \boldsymbol{\alpha}^T \int \boldsymbol{x} \left[\psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{x}) - \mathbb{E}\left\{ \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X}) | \boldsymbol{\alpha}^T \boldsymbol{X} = \boldsymbol{\alpha}^T \boldsymbol{x} \right\} \right] dG(\boldsymbol{x}) \\ &= \mathbb{E}\left[\mathrm{Cov}\left[\boldsymbol{\alpha}^T \boldsymbol{X}, \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X}) | \boldsymbol{\alpha}^T \boldsymbol{X} \right] \right] = \mathbf{0}.\end{aligned}$$

We next note that,

$$(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \bar{\xi}(\boldsymbol{\alpha}) = \mathbb{E}\left[\mathsf{Cov}\left[(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \boldsymbol{X}, \psi_0(\boldsymbol{\alpha}_0^T \boldsymbol{X}) | \boldsymbol{\alpha}^T \boldsymbol{X}\right]\right]$$
$$= \mathbb{E}\left[\mathsf{Cov}\left[(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \boldsymbol{X}, \psi_0(\boldsymbol{\alpha}^T \boldsymbol{X} + (\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \boldsymbol{X}) | \boldsymbol{\alpha}^T \boldsymbol{X}\right]\right],$$

which is positive by the monotonicity of ψ_0. This can be seen as follows. Recall that, using Fubini's theorem, one can prove that for any random variables X and Y such that XY, X and Y are integrable, we have

$$\mathsf{Cov}\{X, Y\} = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y = \int \{\mathbb{P}(X \geq s, Y \geq t) - \mathbb{P}(X \geq s)\mathbb{P}(Y \geq t)\}\, ds\, dt.$$

Denote $Z_1 = (\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \boldsymbol{X}$ and $Z_2 = \psi_0(u + (\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \boldsymbol{X}) = \psi_0(u + Z_1)$, then, using monotonicity of the function ψ_0, we have

$$\mathbb{P}(Z_1 \geq z_1, Z_2 \geq z_2) = \mathbb{P}(Z_1 \geq \max\{z_1, \tilde{z}_2\})$$
$$\geq \mathbb{P}(Z_1 \geq \max\{z_1, \tilde{z}_2\})\mathbb{P}(Z_1 \geq \min\{z_1, \tilde{z}_2\})$$
$$= \mathbb{P}(Z_1 \geq z_1)\mathbb{P}(Z_2 \geq z_2),$$

where

$$\tilde{z}_2 = \psi_0^{-1}(z_2) - u = \inf\{t \in \mathbb{R} : \psi_0(t) \geq z_2\} - u.$$

We conclude that,

$$\mathsf{Cov}\left\{(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \boldsymbol{X}, \psi_0(\boldsymbol{\alpha}^T \boldsymbol{X} + (\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})' \boldsymbol{X}) | \boldsymbol{\alpha}^T \boldsymbol{X} = u\right\}$$
$$= \int \{\mathbb{P}(Z_1 \geq z_1, Z_2 \geq z_2) - \mathbb{P}(Z_1 \geq z_1)\mathbb{P}(Z_2 \geq z_2)\}\, ds\, dt \geq 0,$$

and hence the first part of the Lemma follows. We next prove the uniqueness of the parameter $\boldsymbol{\alpha}_0$. We start by assuming that, on the contrary, there exists $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_0$ in $\mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$ such that

$$(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \bar{\xi}(\boldsymbol{\alpha}) \geq 0 \qquad \text{and } (\boldsymbol{\alpha}_1 - \boldsymbol{\alpha})^T \bar{\xi}(\boldsymbol{\alpha}) \geq 0 \qquad \text{for all } \boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0),$$

and we consider the point $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$ such that

$$|\alpha_j - \alpha_{j0}| = |\alpha_j - \alpha_{j1}| \qquad \text{for } 1 \leq j \leq d.$$

For this point, we have,

$$(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \bar{\xi}(\boldsymbol{\alpha}) = -(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha})^T \bar{\xi}(\boldsymbol{\alpha}) \qquad \text{for all } \boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0),$$

which is not possible since both terms should be positive. This completes the proof of Lemma C.5.2. □

Lemma C.5.3. *Let $f : \mathcal{X} \to \mathbb{R}^k,\ k \leq d$ be a differentiable function on $\mathcal{X}$ such that there exists a constant $M > 0$ satisfying $\|f\|_\infty \leq M$. Then, under the assumptions A1 and A5 we can find a constant $\tilde{M} > 0$ such that for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$ we have that*

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left| \mathbb{E}[f(\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = \boldsymbol{\alpha}^T\boldsymbol{x}] - \mathbb{E}[f(\boldsymbol{X})|\boldsymbol{\alpha}_0^T\boldsymbol{X} = \boldsymbol{\alpha}_0^T\boldsymbol{x}] \right| \leq M\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|.$$

Proof. We can assume without loss of generality that $\alpha_{0,1} \neq 0$ where $\alpha_{0,1}$ is the first component of $\boldsymbol{\alpha}_0$. At the cost of taking a smaller δ_0, we can further assume that $\tilde{\boldsymbol{\alpha}}_1 \neq 0$ for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$. Consider the change of variables $t_1 = \boldsymbol{\alpha}^T\boldsymbol{X}$, $t_i = x_i$ for $1 \leq i \leq d$. Then, the density of $(\boldsymbol{\alpha}^T\boldsymbol{X}, X_2, \ldots, X_d)$ is given by

$$g_{(\boldsymbol{\alpha}^T\boldsymbol{X}, X_2,\ldots,X_d)}(t_1, \ldots, t_d) = g\left(\frac{1}{\alpha_1}\Big(t_1 - \alpha_2 t_2 - \ldots - \alpha_d t_d), t_2, \ldots, t_d\right)\frac{1}{\alpha_1}.$$

Then, for $i = 2, \ldots, d$, the conditional density $h_{\boldsymbol{\alpha}}(x_2, \ldots, x_d|u)$ of the $(d-1)$-dimensional vector $(X_2, \ldots, X_d)$ given that $\boldsymbol{\alpha}^T\boldsymbol{X} = u$ is equal to

$$h_{\boldsymbol{\alpha}}(x_2, \ldots, x_d|u) = \frac{g\left(\frac{u - \alpha_2 x_2 - \ldots - \alpha_d x_d}{\alpha_1}, x_2, \ldots, x_d\right)}{\int g\Big(\frac{u - \alpha_2 t_2 - \ldots - \alpha_d t_d}{\alpha_1}, t_2, \ldots, t_d\Big) \prod_{j=2}^d dt_j} \tag{C.5.5}$$

where the domain of integration in the denominator is the set $\{(x_2, \ldots, x_d) : \boldsymbol{x} \in \mathcal{X})\}$. Note that $X_1 = (\boldsymbol{\alpha}^T\boldsymbol{X} - \alpha_2 X_2 - \ldots - \alpha_d X_d)/\alpha_1$. Thus we have that

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = \boldsymbol{\alpha}^T\boldsymbol{x}] = \mathbb{E}[f(X_1, X_2, \ldots, X_d)|\boldsymbol{\alpha}^T\boldsymbol{X} = \boldsymbol{\alpha}^T\boldsymbol{x}] \\
&= \mathbb{E}\left[f\left(\frac{\boldsymbol{\alpha}^T\boldsymbol{X} - \alpha_2 X_2 - \ldots - \alpha_d X_d}{\alpha_1}, X_2, \ldots, X_d\right) \mid \boldsymbol{\alpha}^T\boldsymbol{X} = \boldsymbol{\alpha}^T\boldsymbol{x}\right] \\
&= \int f\left(\frac{\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2 x_2 - \ldots - \alpha_d x_d}{\alpha_1}, x_2, \ldots, x_d\right) h_{\boldsymbol{\alpha}}(x_2, \ldots, x_d|\boldsymbol{\alpha}^T\boldsymbol{x}) \prod_{j=2}^d dx_j.
\end{aligned}$$

Note now that function

$$\boldsymbol{\alpha} \mapsto h_{\boldsymbol{\alpha}}(x_2, \ldots, x_d|\boldsymbol{\alpha}^T\boldsymbol{x}) = \frac{g\left(\frac{\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2 x_2 - \ldots - \alpha_d x_d}{\alpha_1}, x_2, \ldots, x_d\right)}{\int g\Big(\frac{\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2 t_2 - \ldots - \alpha_d t_d}{\alpha_1}, t_2, \ldots, t_d\Big) \prod_{j=2}^d dt_j},$$

is continuously differentiable on $\mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$. This follows from assumptions A1 and A5 together with Lebesgue dominated convergence theorem which allows us to differentiate the density g under the integral sign. With some notation abuse we write $\partial h/\partial x_i$ for the i-th partial derivative of $\boldsymbol{\alpha} \mapsto h_{\boldsymbol{\alpha}}(x_2, \ldots, x_d|\boldsymbol{\alpha}^T\boldsymbol{x})$. Straightforward calculations yield

$$\begin{aligned}
\frac{\partial h_{\boldsymbol{\alpha}}}{\partial \alpha_1} &= g\left(\frac{\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2 x_2 - \ldots - \alpha_d x_d}{\alpha_1}, x_2, \ldots, x_d\right) \\
&\quad \times \frac{\int \sum_{i=2}^d (x_i - t_i) \frac{\partial g}{\partial x_1}\left(\frac{\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2 t_2 - \ldots - \alpha_d t_d}{\alpha_1}, t_2, \ldots, t_d\right) \prod_{j=2}^d dt_j}{\alpha_1^2 \left(\int g\Big(\frac{\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2 t_2 - \ldots - \alpha_d t_d}{\alpha_1}, t_2, \ldots, t_d\Big) \prod_{j=2}^d dt_j\right)^2},
\end{aligned}$$

and for $2 \le i \le d$

$$\frac{\partial h_{\boldsymbol{\alpha}}}{\partial \alpha_i} = -g\left(\frac{\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2 x_2 - \ldots - \alpha_d x_d}{\alpha_1}, x_2, \ldots, x_d\right)$$
$$\times \frac{\int (x_i - t_i)\frac{\partial g}{\partial x_i}\left(\frac{\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2 t_2 - \ldots - \alpha_d t_d}{\alpha_1}, t_2, \ldots, t_d)\right)\prod_{j=2}^d dt_j}{\alpha_1\left(\int g\left(\frac{\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2 t_2 - \ldots - \alpha_d t_d}{\alpha_1}, t_2, \ldots, t_d\right)\prod_{j=2}^d dt_j\right)^2}.$$

Assumptions A1 and A5 allow us to find a constant $D > 0$ depending on R, $\underline{c}_0$, $\bar{c}_0$ and $\bar{c}_1$ such that

$$\Big\|\frac{\partial h_{\boldsymbol{\alpha}}}{\partial \alpha_i}\Big\|_\infty \le D,$$

for $1 \le i \le d$. Consider now the function $\alpha \mapsto E[f(X)|\boldsymbol{\alpha}^T\boldsymbol{X} = \boldsymbol{\alpha}^T\boldsymbol{x}]$. Using the assumptions of the lemma and applying again Lebesgue dominated convergence theorem we conclude for $i \in \{1, \ldots, d\}$ that we have

$$\frac{\partial \mathbb{E}[f(X)|\boldsymbol{\alpha}^T\boldsymbol{X} = \boldsymbol{\alpha}^T\boldsymbol{x}]}{\partial \alpha_i}$$
$$= \int f\left(\frac{\boldsymbol{\alpha}^T\boldsymbol{x} - \alpha_2 x_2 - \ldots - \alpha_d x_d}{\alpha_1}, x_2, \ldots, x_d\right)\frac{\partial h_{\boldsymbol{\alpha}}(x_2, \ldots, x_d|\boldsymbol{\alpha}^T\boldsymbol{x})}{\partial \alpha_i}\prod_{j=2}^d dx_j.$$

Furthermore, we have that

$$\sup_{(\boldsymbol{x},\mathcal{X}}\Big|\frac{\partial \mathbb{E}[f(\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = \boldsymbol{\alpha}^T\boldsymbol{x}]}{\partial \alpha_i}\Big| \le MD\int\prod_{j=2}^d dx_j = M',$$

for all $i \in \{1, \ldots, d\}$ and $(\boldsymbol{x}, \mathcal{X})$ and $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta)$. The results now follow using a first order Taylor expansion to obtain

$$\Big|\mathbb{E}[f(\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X} = \boldsymbol{\alpha}^T\boldsymbol{x}] - \mathbb{E}[f(\boldsymbol{X})|\boldsymbol{\alpha}_0^T\boldsymbol{X} = \boldsymbol{\alpha}_0^T\boldsymbol{x}]\Big|$$
$$= \Big|\sum_{i=1}^d \frac{\partial \mathbb{E}[f(\boldsymbol{X})|\widetilde{\boldsymbol{\alpha}}^T\boldsymbol{X} = \widetilde{\boldsymbol{\alpha}}^T\boldsymbol{x}]}{\partial \alpha_i}(\alpha_i - \alpha_{0,i})\Big|,$$

for some $\widetilde{\boldsymbol{\alpha}} \in \mathbb{R}^d$ such that $\|\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0\| \le \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|$. Bounding the right side of the preceding display by $\tilde{M}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|$ with $\tilde{M} = dM'$ gives the result.

□

Lemma C.5.4. *Denote for $i \in \{1, \ldots, d\}$ the ith component of the function $u \mapsto \mathbb{E}[\boldsymbol{X}|\boldsymbol{\alpha}^T\boldsymbol{X} = u]$ by $E_{i,\boldsymbol{\alpha}}$. Then $E_{i,\boldsymbol{\alpha}}$ has a total bounded variation. Furthermore, there exists a constant $B > 0$ such that for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$*

$$\|E_{i,\boldsymbol{\alpha}}\|_\infty \le B, \quad \text{and} \quad \int_{\mathcal{I}_{\boldsymbol{\alpha}}} |E'_{i,\boldsymbol{\alpha}}(u)|du \le B,$$

where $\mathcal{I}_{\boldsymbol{\alpha}} = \{\boldsymbol{\alpha}^T\boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}\}$.

Proof. Since $\mathcal{X} \subset \mathcal{B}(0,R)$, it is clear that $\|E_{i,\boldsymbol{\alpha}}\|_\infty \leq R$. As above let us assume without loss of generality that the first component of $\boldsymbol{\alpha}_0$ is not equal to 0. At the cost of taking a smaller δ_0, we can further assume that $\tilde{\alpha}_1 \neq 0$ for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$. We known that for $2 \leq i \leq d$

$$E_{i,\boldsymbol{\alpha}}(u) = \int x_i h_{\boldsymbol{\alpha}}(x_2, \ldots, x_d | u) dx_2 \ldots dx_d,$$

where integration is done over the set $\{(x_2, \ldots, x_d) : (\boldsymbol{x}, \mathcal{X})\}$ and $u \in \mathcal{I}_{\boldsymbol{\alpha}} \subset (a_0 - \delta_0 R, b_0 + \delta_0 R)$ and where $h_{\boldsymbol{\alpha}}$ denotes conditional density of $(X_2, \ldots, X_d)'$ given $\boldsymbol{\alpha}^T \boldsymbol{X} = u$, defined in (C.5.5). Using assumptions A1 and A5 along with the Lebesgue dominated convergence theorem we are allowed to write

$$E'_{i,\boldsymbol{\alpha}}(u) = \int x_i \frac{\partial}{\partial u} h_{\boldsymbol{\alpha}}(x_2, \ldots, x_d | u) \ dx_2 \ldots dx_d.$$

Straightforward calculations yield that

$$\begin{aligned}\frac{\partial}{\partial u} h_{\boldsymbol{\alpha}}(x_2, \ldots, x_d|u) = & \frac{\frac{\partial g}{\partial x_1}\left(\frac{u - \alpha_2 x_2 - \ldots - \alpha_d x_d}{\alpha_1}, x_2, \ldots, x_d\right)}{\alpha_1 \left(\int g\Big(\frac{u-\alpha_2 t_2 - \ldots - \alpha_d t_d}{\alpha_1}, t_2, \ldots, t_d\Big) \prod_{j=2}^d dt_j\right)} \\ & - g\left(\frac{1}{\alpha_1}\Big(u - \alpha_2 x_2 - \ldots - \alpha_d x_d), x_2, \ldots, x_d\right) \\ & \cdot \frac{\int \frac{\partial g}{\partial x_1}\Big(\frac{u-\alpha_2 t_2 - \ldots - \alpha_d t_d}{\alpha_1}, t_2, \ldots, t_d\Big) \prod_{j=2}^d dt_j}{\alpha_1 \left(\int g\Big(\frac{u-\alpha_2 t_2 - \ldots - \alpha_d t_d}{\alpha_1}, t_2, \ldots, t_d\Big) \prod_{j=2}^d dt_j\right)^2}.\end{aligned}$$

Thus, we can find constant $C > 0$ depending only on $|\alpha_{0,1}|$, $\underline{c}_0$, $\underline{c}_1$, $\bar{c}_1$ and R such that $\int |E'_{i,\boldsymbol{\alpha}}(u)| du \leq C$ for all $\alpha \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta_0)$. Now $B = \max(R, C)$ gives the claimed inequalities. If $i = 1$, then

$$E_{1,\boldsymbol{\alpha}}(u) = \frac{1}{\alpha_1}\Big(u - \alpha_j \sum_{j=2}^d E_{j,\boldsymbol{\alpha}}(u)\Big), \ \text{ and } \ e'_{1,\boldsymbol{\alpha}}(u) = \frac{1}{\alpha_1}\Big(1 - \alpha_j \sum_{j=2}^d e'_{j,\boldsymbol{\alpha}}(u)\Big).$$

for $u \in I_{\boldsymbol{\alpha}}$. We conclude again that the claimed inequalities are true at the cost of increasing the constant B obtained above. □

Lemma C.5.5. *Let f be a function defined on some interval $[a, b]$ such that $\|f\|_\infty \leq M$ and*

$$V(f, [a, b]) = \sup_{a = x_0 < x_1 \ldots < x_n = b} \sum_{j=1}^n |f(x_j) - f(x_{j-1})| \leq M,$$

for some finite constant $M > 0$. Then, there exist two nondecreeasing functions f_1 and f_2 on $[a, b]$ such that $\|f_1\|_\infty, \|f_2\|_\infty \leq 2M$ and $f = f_2 - f_1$.

Proof. The fact that $f = f_2 - f_1$ with f_1 and f_2 nondecreeasing on $[a,b]$ follows from the well-known Jordan's decomposition. Furthermore, we can take $f_1(\boldsymbol{x}) = V(f,[a,x])$ and $f_2(\boldsymbol{x}) = f(\boldsymbol{x}) - f_1(\boldsymbol{x})$for $(\boldsymbol{x},[a,b]$. By assumption, $\|f_1\|_\infty \leq M \leq 2M$ and $\|f_2\| \leq \|f\|_\infty + \|f_1\|_\infty \leq 2M.$ □

Lemma C.5.6. *Under Assumptions A4-A5, we can find a constant $C > 0$ such that for all $\boldsymbol{\alpha}$ close enough to $\boldsymbol{\alpha}_0$ we have that*

$$\psi'_{\boldsymbol{\alpha}}(u) > C,$$

for all $u \in \mathcal{I}_{\boldsymbol{\alpha}}$.

Proof. We assume again that $a_1 \neq 0$. By calculations similar to the calculations made in the proof of Lemma C.5.1, we get

$$\begin{aligned}
&\psi_{\boldsymbol{\alpha}}(u)\\
&= \frac{\alpha_{01}}{\alpha_1}\int \psi'_0\left(\frac{\alpha_{01}}{\alpha_1}(u-\alpha_2\tilde{x}_2-\ldots-\alpha_d\tilde{x}_d)+\sum_{j=2}^d \alpha_{0j}\tilde{x}_j\right) h_{\boldsymbol{\alpha}}(\tilde{x}_2,\ldots,\tilde{x}_d|u)\prod_{j=2}^d d\tilde{x}_j\\
&\quad+\int \psi_0\left(\frac{\alpha_{01}}{\alpha_1}(u-\alpha_2\tilde{x}_2-\ldots-\alpha_d\tilde{x}_d)+\sum_{j=2}^d \alpha_{0j}\tilde{x}_j\right)\frac{\partial}{\partial u}h(\tilde{x}_2,\ldots,\tilde{x}_d|u)\prod_{j=2}^d d\tilde{x}_j.
\end{aligned}$$

Now, a Taylor expansion of α_i in the neighborhood of $\alpha_{0,i}$ and using that $\alpha_{0,1}/\alpha_1 = 1 - \epsilon_1/\alpha_{0,1} + o(\epsilon_1)$ yields

$$\begin{aligned}
&\psi_0\left(\frac{\alpha_{0,1}}{\alpha_1}(u-\alpha_2x_2-\cdots-\alpha_dx_d)+\alpha_{0,2}x_2+\cdots+\alpha_{0,d}x_d\right)\\
&= \psi_0\left(u-\frac{\epsilon_1}{\alpha_{0,1}}(u-\epsilon_2x_2-\ldots-\epsilon_dx_d)+o(\epsilon_1)\right)\\
&= \psi_0(u)-\frac{\epsilon_1}{\alpha_{0,1}}(u-\epsilon_2x_2-\ldots-\epsilon_dx_d)\psi'_0(u)+o(\epsilon_1)\\
&= \psi_0(u)-\frac{\epsilon_1}{\alpha_{0,1}}u\psi'_0(u)+o(\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|).
\end{aligned}$$

Using the Lebesgue dominated convergence theorem and the fact that $h_{\boldsymbol{\alpha}}(\tilde{x}_2,\ldots,\tilde{x}_d|u)$ is a conditional density it follows that

$$\begin{aligned}
&\int \psi_0\left(\frac{\alpha_{01}}{\alpha_1}(u-\alpha_2\tilde{x}_2-\ldots-\alpha_d\tilde{x}_d)+\sum_{j=2}^d \alpha_{0j}\tilde{x}_j\right)\frac{\partial}{\partial u}h(\tilde{x}_2,\ldots,\tilde{x}_d|u)\prod_{j=2}^d d\tilde{x}_j,\\
&= o(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0),
\end{aligned}$$

such that

$$\psi'_{\boldsymbol{\alpha}}(u) \geq C\left(1 - \frac{\epsilon_1}{\alpha_{0,1}}\right) + o(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) \geq C > 0,$$

provided that $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\|$ is small enough. □

Lemma C.5.7. *If $h \asymp n^{-1/7}$, then there exists a constant $B > 0$ such that for all $\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta)$*

$$\|\psi'_{nh,\boldsymbol{\alpha}}\|_\infty \leq B \qquad \text{and} \qquad \int_{\mathcal{I}_{\boldsymbol{\alpha}}} |\psi''_{nh,\boldsymbol{\alpha}}(u)| du \leq B,$$

where $\mathcal{I}_{\boldsymbol{\alpha}} = \{\boldsymbol{\alpha}^T \boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}\}$

Proof. Using integration by parts and Proposition 4.2.2, we have for all $u \in \mathcal{I}_{\boldsymbol{\alpha}}$

$$\begin{aligned}
\psi'_{nh,\boldsymbol{\alpha}}(u) &= \frac{1}{h}\int K\left(\frac{u-x}{h}\right) d\hat{\psi}_{n\boldsymbol{\alpha}}(x) \\
&= \frac{1}{h}\int K\left(\frac{u-x}{h}\right) \psi'_{\boldsymbol{\alpha}}(x)dx + \frac{1}{h^2}\int K'\left(\frac{u-x}{h}\right)(\hat{\psi}_{n\boldsymbol{\alpha}}(x) - \psi_{\boldsymbol{\alpha}}(x))dx \\
&= \frac{1}{h}\int K\left(\frac{u-x}{h}\right) d\psi_{\boldsymbol{\alpha}}(x) + \frac{1}{h}\int K'(w)(\hat{\psi}_{n\boldsymbol{\alpha}}(u+hw) - \psi_{\boldsymbol{\alpha}}(u+hw))dw \\
&= \psi'_{\boldsymbol{\alpha}}(u) + O(h^2) + O_p(h^{-1}\log n n^{-1/3}) = \psi'_{\boldsymbol{\alpha}}(u) + o_p(1).
\end{aligned}$$

This proves the first part of Lemma C.5.7. For the second part, we get by a similar calculation that,

$$\begin{aligned}
\psi''_{nh,\boldsymbol{\alpha}}(u) &= \frac{1}{h}\int K\left(\frac{u-x}{h}\right) \psi''_{\boldsymbol{\alpha}}(x)dx + \frac{1}{h^2}\int K''(w)(\hat{\psi}_{n\boldsymbol{\alpha}}(u+hw) - \psi_{\boldsymbol{\alpha}}(u+hw))dw \\
&= \frac{1}{h}\int K\left(\frac{u-x}{h}\right) \psi''_{\boldsymbol{\alpha}}(x)dx + O_p(h^{-2}\log n n^{-1/3}).
\end{aligned}$$

Since $h^{-2}\log n n^{-1/3} = o(1)$ for $h \asymp n^{-1/7}$, the second part follows by Assumption A10. □

Appendix D

Varying coefficient models - Appendix

D.1 Definitions and properties

This section contains the Definition of the L_2-distance and the Assumptions needed for the main results, i.e., Theorem 5.5.1 and 5.5.2.

Assumption A.

1. *For all $1 \leq p \leq d$, the random variable U_p has distribution function F_{U_p} on $\mathcal{U}_p = [a_p, b_p]$. The distribution function F_{U_p} has Lebesgue density f_{U_p} which is bounded away from zero and infinity, uniformly in $\mathcal{U}_p$, i.e. there exist positive constants N_1 and N_2 such that $N_1 \leq f_{U_p}(u) \leq N_2$ for $u \in \mathcal{U}_p$.*

2. *The eigenvalues $\eta_1(\mathbf{u}), \ldots, \eta_d(\mathbf{u})$ of $\Sigma(\mathbf{u}) = \mathbb{E}(\mathbf{X}\mathbf{X}^T|\mathbf{U} = \mathbf{u})$ are bounded away from zero and infinity, uniformly over all $\mathbf{u} \in \mathcal{U}^d$, i.e. there exist positive constants N_3 and N_4 such that $N_3 \leq \eta_1(\mathbf{u}) \leq \cdots \leq \eta_d(\mathbf{u}) \leq N_4$ for $\mathbf{u} \in \mathcal{U}^d$.*

3. *There exists a positive constant N_5 such that $\mid X_p \mid \leq N_5$ for $1 \leq p \leq d$.*

4. *There exists a positive constant N_6 such that $\sigma_j^2(\mathbf{u}, \mathbf{x}) \leq N_6 < \infty$ for $j = 1, 2$ and for every $\mathbf{u} \in \mathcal{U}^d, \mathbf{x} \in \mathbb{R}^d$, where $\sigma_j^2(\mathbf{u}, \mathbf{x}) = \mathrm{Var}(Y_j^* \mid \mathbf{U} = \mathbf{u}, \mathbf{X} = \mathbf{x})$.*

5. $\limsup_n \left(\frac{\max_p m_p}{\min_p m_p}\right) < \infty$.

6. $n^{-1} m_{max}^{3/2} \lambda_{max} \to 0$ *and* $n^{-1} m_{max} \to 0$ *as* $n \to \infty$.

7. $n^{-1}m_{max}\log(m_{max}) \to 0$ *as* $n \to \infty$.

8. $\rho_n \to 0$ *as* $n \to \infty$.

Assumption B.

1. $\sup_{\mathbf{u},\mathbf{x}} \tau_1(\mathbf{u},\mathbf{x}) \sup_{t\le\tau_1(\mathbf{u},\mathbf{x})} |\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| = o_p(1)$.

2. $\sup_{\mathbf{u},\mathbf{x}} \kappa(\mathbf{u},\mathbf{x}) \to 0$ *as* $n \to \infty$.

3. $\sup_{\mathbf{u},\mathbf{x}} \kappa_\sigma(\mathbf{u},\mathbf{x}) \to 0$ *as* $n \to \infty$.

Assumption C.

1. $\beta_p \in C^3([a_p,b_p])$, *for each* $1 \le p \le d$, *where* $C^r([a,b])$ *is the space of* $r-$*times continuously differentiable functions on* $[a,b]$.

2. $m_{max}^{3/2}\Big[\sup_{\mathbf{u},\mathbf{x}}\{\tau_1(\mathbf{u},\mathbf{x}) \sup_{t\le\tau_1(\mathbf{u},\mathbf{x})} |\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + \kappa(\mathbf{u},\mathbf{x})$
$+ \kappa_\sigma(\mathbf{u},\mathbf{x})\} + \rho_n\Big] \to 0$ *and* $n^{-1/2}m_{max}^2 \to 0$; $n^{-1}m_{max}^{3/2}\lambda_{max} \to 0$
as $n \to \infty$.

Assumption D.

1. $m_{max}^{-1/2}n^{1/2}\Big(\sup_{\mathbf{u},\mathbf{x}}\big(\tau_1(\mathbf{u},\mathbf{x}) \sup_{t\le\tau_1(\mathbf{u},\mathbf{x})} |\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| +$
$\kappa(\mathbf{u},\mathbf{x})\big)\Big) \to 0$ *and* $n^{-1/2}m_{max}\lambda_{max} + n^{1/2}\rho_n \to 0$ *as* $n \to \infty$.

2. $m_{max}^{-1}n^{1/2} \sup_{\mathbf{u},\mathbf{x}}\big(\tau_1(\mathbf{u},\mathbf{x}) \sup_{t\le\tau_1(\mathbf{u},\mathbf{x})} |\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + \kappa(\mathbf{u},\mathbf{x})$
$+ \quad \kappa_\sigma(\mathbf{u},\mathbf{x})\big) \to 0$ *and* $m_{max}^{-1/2}\log n + n^{-1/2}m_{max}\lambda_{max} + n^{1/2}\rho_n \to 0$
as $n \to \infty$.

Assumption A.1 guarantees that the observation points are randomly scattered and is a natural assumption in nonparametric regression (see e.g. Eubank and Speckman (1990)). All A assumptions are common in P-spline theory (see e.g. Antoniadis et al. (2012)). In particular A.1-A.4 are common in mean regression in varying coefficient models. Also note that Assumptions A.5, A.6 andA.7 are satisfied with the choice of number of knots and smoothing parameter of Remark 5.5.2. When all β_p have bounded r-th derivatives $\rho_n = O_p(m_{\max}^{-r})$ (Schumaker (2007)). Assumption B ensures that the censored nature of the data is taken into account and is illustrated by an example in Remark 5.5.1. When the Kaplan-Meier estimator is used to estimate G, it follows from Zhou (1991) that $\sup_{t\le\tau_1(\mathbf{u},\mathbf{x})} |\hat{G}(t) - G(t)| = O_p(n^{-1/2})$. Assumption C guarantees that, uniformly over $\mathcal{U}_p$, the second order derivative of $\hat{\beta}_{1p}$ is a consistent estimator for β_{1p}, for $1 \le p \le d$. It is a technical assumption needed in the proof of Theorem 5.5.1, Part 2 and guarantees that the Kaplan-Meier estimator based on residual observations constructed with method

1 converges to the true error distribution F. Assumption D is an assumption on the convergence rate of the P-spline estimators and guarantees that the squared L_∞-distance between the P-spline estimators $\hat{\boldsymbol{\beta}}_j$ and $\boldsymbol{\beta}_j^*$ converges to zero at a faster rate than the variance given by (5.5.1). For the examples considered in Remark 5.5.1, Assumptions C and D are also fulfilled when G is estimated using the Kaplan-Meier estimator, $m_{\mathsf{max}} \asymp n^{1/5}$ and $\lambda_{\mathsf{max}} \asymp n^{\iota}, \iota < 3/10$.

Definition D.1.1. *For a real valued function f on $\mathcal{U}$ and a vector valued function $\mathbf{g} = (g_1, ..., g_d)$ on $\mathcal{U}^d$, the L_2-norm is given by:*

$$\|f\|_{L_2} = \left\{ \int_{\mathcal{U}} f^2(t)dt \right\}^{1/2}, \quad \|\mathbf{g}\|_{L_2} = \left(\sum_{p=1}^{d} \|g_p\|_{L_2}^2 \right)^{1/2},$$

Definition D.1.2. *For a real valued matrix $\mathbf{A}$ of dimension $m_A \times n_A$, the 2-norm of $\mathbf{A}$ is given by $\|\mathbf{A}\|_2 = \sup_{\mathbf{x} \neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$, with $\mathbf{x} \in \mathbb{R}^{n_A}$ and $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n_A} x_i^2}$. This norm is equal to $\sqrt{\zeta_{max}(\mathbf{A}^T\mathbf{A})}$ where ζ_{max} is the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$.*

Definition D.1.3. *For sequences of positive numbers r_n and s_n, $r_n \lesssim s_n$ means that $s_n^{-1} r_n$ is bounded and $r_n \asymp s_n$ means that $s_n^{-1} r_n$ and $r_n^{-1} s_n$ are bounded.*

Definition D.1.4. *For a real valued function f on $\mathcal{U}$ and a vector valued function $\mathbf{g} = (g_1, ..., g_d)$ on $\mathcal{U}^d$, the L_∞-norm is given by:*

$$\|f\|_\infty = \sup_{u \in \mathcal{U}} |\, f(u) \,|, \quad \|\mathbf{g}\|_\infty = \max_{1 \le p \le d} \|g_p\|_\infty.$$

Our estimation technique relies on properties of B-splines. For a detailed description of B-splines we refer to De Boor (1978) or Schumaker (2007).

Property D.1.1. $B_{pl}(u_p; q_p) \ge 0; \quad \sum_{l=1}^{m_p} B_{pl}(u_p; q_p) = 1.$

Property D.1.2. *There exists positive constants N_7, N_8 and coefficients $\alpha_{pl} \in \mathbb{R}$ such that*

$$m_p^{-1} N_7 \sum_{l=1}^{m_p} \alpha_{pl}^2 \le \int_{\mathcal{U}} \{ \sum_{l=1}^{m_p} \alpha_{pl} B_{pl}(u_p; q_p) \}^2 du \le m_p^{-1} N_8 \sum_{l=1}^{m_p} \alpha_{pl}^2.$$

Property D.1.3. $\int_{\mathcal{U}} B_{pl}(u; q_p) du = O(m_p^{-1}).$

Property D.1.4. $\|g\|_\infty \lesssim m_p^{-1/2} \|g\|_{L_2}$ *for $g \in \mathbb{G}_p, 1 \le p \le d$ where $\mathbb{G}_p$ is the space of spline functions of degree q_p on $\mathcal{U}_p$ with knots $\boldsymbol{\xi}_p$.*

We use as notations $\hat{\boldsymbol{\alpha}}_j, \boldsymbol{\alpha}_j^*$ and $\tilde{\boldsymbol{\alpha}}_j$ for methods $j = 1, 2$ (described in Section 5.4), when we replace $\mathbf{Y}$ in expression

$$\hat{\boldsymbol{\alpha}} = \left(\mathbf{R}^T \mathbf{R} + \mathbf{Q}_{\boldsymbol{\lambda}} \right)^{-1} \mathbf{R}^T \mathbf{Y}.$$

by $\hat{\mathbf{Y}}_j^* = (\hat{Y}_{j1}^*, \ldots, \hat{Y}_{jn}^*)^T$, $\mathbf{Y}_j^* = (Y_{j1}^*, \ldots, Y_{jn}^*)^T$, and $\mathbf{M} = (M_{j1}, \ldots, M_{jn})^T$ with $M_{ji} = \mathbb{E}(Y_{ji}^*|\mathbf{U}_i, X_i)$ for $1 \leq i \leq n$ respectively. Similar notations hold for $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jd})^T$, $\boldsymbol{\beta}_j^* = (\beta_{j1}^*, \ldots, \beta_{jd}^*)^T$ and $\tilde{\boldsymbol{\beta}}_j = (\tilde{\beta}_{j1}, \ldots, \tilde{\beta}_{jd})^T$.

D.2 Proof of Theorem 5.5.1, Part 1

The proof of the first result stated in Theorem 5.5.1 relies on the maximal distance between the Y_{1i}^* and $\hat{Y}_{1i}^*$ responses, derived in Lemma D.2.1.

Lemma D.2.1. $\max_{1\leq i\leq n} \mid \hat{Y}_{1i}^* - Y_{1i}^* \mid=$

$$O_p\left(\sup_{\mathbf{u},\mathbf{x}}\left\{\tau_1(\mathbf{u},\mathbf{x})\sup_{t\leq\tau_1(\mathbf{u},\mathbf{x})}|\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + \kappa(\mathbf{u},\mathbf{x})\right\}\right),$$

Proof of Lemma D.2.1. Since$\mid \hat{Y}_{1i}^* - Y_{1i}^* \mid=$

$$\mid \hat{Y}_{1i}^* - Y_{1i}^* \mid 1_{\{Z_i\leq\tau_1(\mathbf{U}_i,\mathbf{X}_i)\}} + \mid \hat{Y}_{1i}^* - Y_{1i}^* \mid 1_{\{Z_i>\tau_1(\mathbf{U}_i,\mathbf{X}_i)\}},$$

we consider two cases and prove the following results,

$$\max_{1\leq i\leq n}\left\{\mid \hat{Y}_{1i}^* - Y_{1i}^* \mid 1_{\{Z_i\leq\tau_1(\mathbf{U}_i,\mathbf{X}_i)\}}\right\} \lesssim \sup_{\mathbf{u},\mathbf{x}}\left(\tau_1(\mathbf{u},\mathbf{x})\sup_{t\leq\tau_1(\mathbf{u},\mathbf{x})}|\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})|\right). \quad \text{(D.2.1)}$$

$$\max_{1\leq i\leq n}\left\{\mid \hat{Y}_{1i}^* - Y_{1i}^* \mid 1_{\{Z_i>\tau_1(\mathbf{U}_i,\mathbf{X}_i)\}}\right\} \lesssim \sup_{\mathbf{u},\mathbf{x}}\kappa(\mathbf{u},\mathbf{x}). \quad \text{(D.2.2)}$$

For (D.2.1) we start by the triangle inequality,

$$\begin{aligned}\mid \hat{Y}_{1i}^* - Y_{1i}^* \mid 1_{\{Z_i\leq\tau_1(\mathbf{U}_i,\mathbf{X}_i)\}} &\leq\mid \Delta_i\{\hat{\varphi}_1(\mathbf{U}_i,\mathbf{X}_i,Z_i) - \varphi_1(\mathbf{U}_i,\mathbf{X}_i,Z_i)\} \\ &\quad + (1-\Delta_i)\{\hat{\psi}_1(\mathbf{U}_i,\mathbf{X}_i,Z_i) - \psi_1(\mathbf{U}_i,\mathbf{X}_i,Z_i)\} \mid \\ \leq\mid \hat{\varphi}_1(\mathbf{U}_i,\mathbf{X}_i,Z_i) - \varphi_1(\mathbf{U}_i,\mathbf{X}_i,Z_i) \mid &+ \mid \hat{\psi}_1(\mathbf{U}_i,\mathbf{X}_i,Z_i) - \psi_1(\mathbf{U}_i,\mathbf{X}_i,Z_i) \mid.\end{aligned}$$

We derive the order bound for $\mid \hat{\varphi}_1(\mathbf{U}_i,\mathbf{X}_i,Z_i) - \varphi_1(\mathbf{U}_i,\mathbf{X}_i,Z_i) \mid$, similar result holds if we replace φ_1 and $\hat{\varphi}_1$ by ψ_1 and $\hat{\psi}_1$ respectively.

$$\begin{aligned}
&|\hat{\varphi}_1(\mathbf{U}_i, \mathbf{X}_i, Z_i) - \varphi_1(\mathbf{U}_i, \mathbf{X}_i, Z_i)| \\
&\quad \leq \left|(1+\gamma)\left\{\int_0^{Z_i} \frac{1}{\hat{G}(t|\mathbf{U}_i, \mathbf{X}_i)} dt - \int_0^{Z_i} \frac{1}{G(t|\mathbf{U}_i, \mathbf{X}_i)} dt\right\}\right| \\
&\qquad + \left|\frac{\gamma Z_i}{\hat{G}(Z_i|\mathbf{U}_i, \mathbf{X}_i)} - \frac{\gamma Z_i}{G(Z_i|\mathbf{U}_i, \mathbf{X}_i)}\right| \\
&\quad \leq \left|(1+\gamma)\int_0^{Z_i} \frac{\hat{G}(t|\mathbf{U}_i, \mathbf{X}_i) - G(t|\mathbf{U}_i, \mathbf{X}_i)}{G(t|\mathbf{U}_i, \mathbf{X}_i)\hat{G}(t|\mathbf{U}_i, \mathbf{X}_i)} dt\right| \\
&\qquad + \left|\frac{\gamma Z_i\{\hat{G}(Z_i|\mathbf{U}_i, \mathbf{X}_i) - G(Z_i|\mathbf{U}_i, \mathbf{X}_i)\}}{G(Z_i|\mathbf{U}_i, \mathbf{X}_i)\hat{G}(Z_i|\mathbf{U}_i, \mathbf{X}_i)}\right|
\end{aligned}$$

$$\begin{aligned}
&\leq |1+\gamma| \sup_{t \leq \tau_1(\mathbf{U}_i, \mathbf{X}_i)} \Big\{ |\ \hat{G}(t|\mathbf{U}_i, \mathbf{X}_i) - G(t|\mathbf{U}_i, \mathbf{X}_i)\ | \Big\} \\
&\qquad\qquad \times \int_0^{\tau_1(\mathbf{U}_i, \mathbf{X}_i)} \frac{G(t|\mathbf{U}_i, \mathbf{X}_i)}{\hat{G}(t|\mathbf{U}_i, \mathbf{X}_i)} \frac{1}{G(t|\mathbf{U}_i, \mathbf{X}_i)^2} dt \\
&\quad + |\gamma| \tau_1(\mathbf{U}_i, \mathbf{X}_i) \sup_{t \leq \tau_1(\mathbf{U}_i, \mathbf{X}_i)} \Big\{ |\ \hat{G}(t|\mathbf{U}_i, \mathbf{X}_i) - G(t|\mathbf{U}_i, \mathbf{X}_i)\ | \Big\} \\
&\qquad\qquad \times \sup_{t \leq \tau_1(\mathbf{U}_i, \mathbf{X}_i)} \Big\{ \frac{1}{G(t|\mathbf{U}_i, \mathbf{X}_i)^2} \frac{G(t|\mathbf{U}_i, \mathbf{X}_i)}{\hat{G}(t|\mathbf{U}_i, \mathbf{X}_i)} \Big\}.
\end{aligned}$$

From the uniform convergence of $\hat{G}$ we have:

$$\sup_{t \leq \tau_1(\mathbf{U}_i, \mathbf{X}_i)} \frac{G(t|\mathbf{U}_i, \mathbf{X}_i)}{\hat{G}(t|\mathbf{U}_i, \mathbf{X}_i)} = 1 + o_p(1).$$

Also $\inf_{t \leq \tau_1(\mathbf{U}_i, \mathbf{X}_i)}\{G(t|\mathbf{U}_i, \mathbf{X}_i)\} > 0$, therefore,

$$\begin{aligned}
&|\ \hat{\varphi}_1(\mathbf{U}_i, \mathbf{X}_i, Z_i) - \varphi_1(\mathbf{U}_i, \mathbf{X}_i, Z_i)\ | \\
&\qquad = O_p\Big(\tau_1(\mathbf{U}_i, \mathbf{X}_i) \sup_{t \leq \tau_1(\mathbf{U}_i, \mathbf{X}_i)} |\ \hat{G}(t|\mathbf{U}_i, \mathbf{X}_i) - G(t|\mathbf{U}_i, \mathbf{X}_i)\ | \Big).
\end{aligned}$$

For (D.2.2) we have,

$$\begin{aligned}
&\mathbb{E}\{|\ \hat{Y}^*_{1i} - Y^*_{1i}\ |\ 1_{\{Z_i > \tau_1(\mathbf{U}_i, \mathbf{X}_i)\}}\} \\
&\quad \leq \mathbb{E}\left[\mathbb{E}\left\{\max_{\phi=\varphi_1,\psi_1} 1_{\{Z_i > \tau_1(\mathbf{U}_i, \mathbf{X}_i)\}}\ |\ Z_i - \phi(U_i, \mathbf{X}_i, Z_i)\ |\ \Big| \mathbf{U}_i, \mathbf{X}_i\right\}\right] \\
&\quad \leq \sup_{\mathbf{u},\mathbf{x}} \kappa(\mathbf{u}, \mathbf{x}).
\end{aligned}$$

By combining (D.2.1) and (D.2.2), the result of Lemma D.2.1 follows. □

Proof of Theorem 5.5.1, Part 1. Since

$$\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_{L_2} \leq \|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_{L_2} + \|\boldsymbol{\beta}_1^* - \tilde{\boldsymbol{\beta}}_1\|_{L_2} + \|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_{L_2},$$

the result follows by showing that

$$\begin{aligned} &\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_{L_2} \\ &\quad = O_p\left(\sup_{\mathbf{u},\mathbf{x}}\left\{\tau_1(\mathbf{u},\mathbf{x})\sup_{t\leq\tau_1(\mathbf{u},\mathbf{x})}|\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + \kappa(\mathbf{u},\mathbf{x})\right\}\right) && \text{(D.2.3)} \\ &\|\boldsymbol{\beta}_1^* - \tilde{\boldsymbol{\beta}}_1\|_{L_2} = O_p\left(n^{-1/2}m_{\mathsf{max}}^{1/2}\right) && \text{(D.2.4)} \\ &\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\|_{L_2} = O_p\left(n^{-1}m_{\mathsf{max}}^{3/2}\lambda_{\mathsf{max}} + \rho_n\right). && \text{(D.2.5)} \end{aligned}$$

We start with the proof of (D.2.3). By Property D.1.2 it suffices to show that

$$\begin{aligned} &\|\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^*\|_2 = \\ &O_p\left(m_{\mathsf{max}}^{1/2}\left(\sup_{\mathbf{u},\mathbf{x}}\left\{\tau_1(\mathbf{u},\mathbf{x})\sup_{t\leq\tau_1(\mathbf{u},\mathbf{x})}|\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + \kappa(\mathbf{u},\mathbf{x})\right\}\right)\right). \end{aligned}$$

From Antoniadis et al. (2012) we have,

$$\begin{aligned} &\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^* \\ &= \left\{(\mathbf{R}^T\mathbf{R})^{-1} - (\mathbf{R}^T\mathbf{R})^{-1}\mathbf{Q}_{\boldsymbol{\lambda}}(\mathbf{R}^T\mathbf{R})^{-1} + o_p(n^{-1}m_{\mathsf{max}}^{3/2}\lambda_{\mathsf{max}})(\mathbf{R}^T\mathbf{R})^{-1}\right\} \\ &\qquad\qquad \times\sum_{i=1}^n \mathbf{R}_i(\hat{Y}_{1i}^* - Y_{1i}^*) \\ &= \hat{\boldsymbol{\alpha}}_{1,reg} - \boldsymbol{\alpha}_{reg}^* - \left\{(\mathbf{R}^T\mathbf{R})^{-1}\mathbf{Q}_{\boldsymbol{\lambda}}(\mathbf{R}^T\mathbf{R})^{-1} + o_p(n^{-1}m_{\mathsf{max}}^{3/2}\lambda_{\mathsf{max}})(\mathbf{R}^T\mathbf{R})^{-1}\right\} \\ &\qquad\qquad \times\sum_{i=1}^n \mathbf{R}_i(\hat{Y}_{1i}^* - Y_{1i}^*) \\ &= \left\{1 - (\mathbf{R}^T\mathbf{R})^{-1}\mathbf{Q}_{\boldsymbol{\lambda}} + o_p(n^{-1}m_{\mathsf{max}}^{3/2}\lambda_{\mathsf{max}})\right\}\left(\hat{\boldsymbol{\alpha}}_{1,reg} - \boldsymbol{\alpha}_{reg}^*\right), \end{aligned}$$

where $\hat{\boldsymbol{\alpha}}_{1,reg}$ and $\boldsymbol{\alpha}_{reg}^*$ denote the regular B-spline estimator (i.e. $\lambda_0 = \ldots = \lambda_d = 0$). Consequently,

$$\begin{aligned} &\|\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^*\|_2 \\ &\quad \leq \left\{1 + \|(\mathbf{R}^T\mathbf{R})^{-1}\|_2\|\mathbf{Q}_{\boldsymbol{\lambda}}\|_2 + o_p(n^{-1}m_{\mathsf{max}}^{3/2}\lambda_{\mathsf{max}})\right\}\|\hat{\boldsymbol{\alpha}}_{1,reg} - \boldsymbol{\alpha}_{1,reg}^*\|_2. \end{aligned}$$

From Lemma 1 in Antoniadis et al. (2012) we know that except on an event whose probability tends to zero, $\|(\mathbf{R}^T\mathbf{R})^{-1}\|_2\|\mathbf{Q}_{\boldsymbol{\lambda}}\|_2 = O_p(n^{-1}m_{\mathsf{max}}^{3/2}\lambda_{\mathsf{max}})$,

$$\begin{aligned} &\|\hat{\boldsymbol{\alpha}}_{1,reg} - \boldsymbol{\alpha}_{1,reg}^*\|_2^2 = (\hat{\mathbf{Y}}_1^* - \mathbf{Y}_1^*)^T\mathbf{R}(\mathbf{R}^T\mathbf{R})^{-1}(\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T(\hat{\mathbf{Y}}_1^* - \mathbf{Y}_1^*) \\ &= (n^{-1}m_{\mathsf{max}})^2(\hat{\mathbf{Y}}_1^* - \mathbf{Y}_1^*)^T\mathbf{R}\left(n^{-1}m_{\mathsf{max}}\mathbf{R}^T\mathbf{R}\right)^{-1}\left(n^{-1}m_{\mathsf{max}}\mathbf{R}^T\mathbf{R}\right)^{-1}\mathbf{R}^T(\hat{\mathbf{Y}}_1^* - \mathbf{Y}_1^*). \end{aligned}$$

and since all eigenvalues of $n^{-1}m_{\max}\mathbf{R}^T\mathbf{R}$ fall between positive constants, we have $\|n^{-1}m_{\max}\mathbf{R}^T\mathbf{R}\|_2 \asymp 1$ and thus,

$$\begin{aligned}&\|\hat{\boldsymbol{\alpha}}_{1,reg} - \boldsymbol{\alpha}^*_{1,reg}\|_2^2 = (\hat{\mathbf{Y}}^*_1 - \hat{\mathbf{Y}}^*_1)^T\mathbf{R}(\mathbf{R}^T\mathbf{R})^{-1}(\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T(\hat{\mathbf{Y}}^*_1 - \mathbf{Y}^*_1)\\ &\asymp n^{-1}m_{\max}(\hat{\mathbf{Y}}^*_1 - \mathbf{Y}^*_1)^T(\hat{\mathbf{Y}}^*_1 - \mathbf{Y}^*_1)\\ &\lesssim m_{\max}\left(\sup_{\mathbf{u},\mathbf{x}}\left\{\tau_1(\mathbf{u},\mathbf{x})\sup_{t\le\tau_1(\mathbf{u},\mathbf{x})}|\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + \kappa(\mathbf{u},\mathbf{x})\right\}\right)^2.\end{aligned}$$

In the last step, we use the result of Lemma D.2.1 and the inequality

$$\sqrt{(\hat{\mathbf{Y}}^*_1 - \mathbf{Y}^*_1)^T(\hat{\mathbf{Y}}^*_1 - \mathbf{Y}^*_1)} = \|\hat{\mathbf{Y}}^*_1 - \mathbf{Y}^*_1\|_2 \le \sqrt{n}\max_{1\le i\le n}|\hat{Y}^*_{1i} - Y^*_{1i}|.$$

We continue with the proof of (D.2.4). Using similar arguments as is the proof of (D.2.3), we have

$$\begin{aligned}&\|\boldsymbol{\alpha}^*_1 - \tilde{\boldsymbol{\alpha}}_1\|_2\\ &\le \left\{1 + \|(\mathbf{R}^T\mathbf{R})^{-1}\|_2\|\mathbf{Q}_{\boldsymbol{\lambda}}\|_2 + o_p(n^{-1}m_{\max}^{3/2}\lambda_{\max})\right\}\|\boldsymbol{\alpha}^*_{1,reg} - \tilde{\boldsymbol{\alpha}}_{1,reg}\|_2, \qquad \text{(D.2.6)}\end{aligned}$$

and,

$$\begin{aligned}&\|\boldsymbol{\alpha}^*_{1,reg} - \tilde{\boldsymbol{\alpha}}_{1,reg}\|_2^2\\ &= (n^{-1}m_{\max})^2(\mathbf{Y}^*_1 - \mathbf{M}_1)^T\mathbf{R}(n^{-1}m_{\max}\mathbf{R}^T\mathbf{R})^{-1}(n^{-1}m_{\max}\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T(\mathbf{Y}^*_1 - \mathbf{M}_1).\end{aligned}$$

By Assumption A.3,

$$\begin{aligned}&\mathbb{E}\left\{(\mathbf{Y}^*_1 - \mathbf{M}_1)^T\mathbf{R}\mathbf{R}^T(\mathbf{Y}^*_1 - \mathbf{M}_1)\right\}\\ &= \mathbb{E}\left[\left\{\sum_{i=1}^n\mathbf{R}_i(Y^*_{1i} - M_{1i})\right\}^T\left\{(\sum_{i=1}^n\mathbf{R}_i(Y^*_{1i} - M_{1i})\right\}\right]\\ &= \mathbb{E}\left\{\sum_{p,l}\sum_{i,j=1}^n X_{ip}X_{jp}B_{pl}(U_{ip};q_p)B_{pl}(U_{jp};q_p)(Y^*_{1i} - M_{1i})(Y^*_{1j} - M_{1j})\right\}\\ &\lesssim \sum_{p,l}\left[\sum_{i=1}^n\mathbb{E}\left\{B^2_{pl}(U_{ip};q_p)^2(Y^*_{1i} - M_{1i})^2\right\}\right.\\ &\qquad\left.+\sum_{i\ne j}\mathbb{E}\left\{B_{pl}(U_{ip};q_p)B_{pl}(U_{jp};q_p)(Y^*_{1i} - M_{1i})(Y^*_{1j} - M_{1j})\right\}\right].\end{aligned}$$

By the independence of the observations, Assumption A.5 and Properties D.1.2 and D.1.3 of B-splines it follows that, using the law of the total expectation,

$$\begin{aligned}&\mathbb{E}\left\{B^2_{pl}(U_{ip};q_p)(Y^*_{1i} - M_{1i})^2\right\} \lesssim \mathbb{E}\{B^2_{pl}(U_{ip};q_p)\} \lesssim m_p^{-1} = O(m_{\max}^{-1}),\\ &\mathbb{E}\{B_{pl}(U_{ip};q_p)B_{pl}(U_{jp};q_p)(Y^*_{1i} - M_{1i})(Y^*_{1j} - M_{1j})\}\\ &\quad = \mathbb{E}\{B_{pl}(U_{ip};q_p)(Y^*_{1i} - M_{1i})\}\mathbb{E}\{B_{pl}(U_{jp};q_p)(Y^*_{1j} - M_{1j})\} = 0.\end{aligned}$$

Therefore,

$$\mathbb{E}\left\{(\mathbf{Y}_1^* - \mathbf{M}_1)^T \mathbf{R}\mathbf{R}^T (\mathbf{Y}_1^* - \mathbf{M}_1)\right\} = O(n),$$
$$(\mathbf{Y}_1^* - \mathbf{M}_1)^T \mathbf{R}\mathbf{R}^T (\mathbf{Y}_1^* - \mathbf{M}_1) = O_p(n),$$

such that

$$\|\boldsymbol{\alpha}_{1,reg}^* - \tilde{\boldsymbol{\alpha}}_{1,reg}\|_2^2 = O_p\left(n^{-1} m_{\max}^2\right). \tag{D.2.7}$$

Combining (D.2.6) and (D.2.7) gives,

$$\|\boldsymbol{\alpha}_1^* - \tilde{\boldsymbol{\alpha}}_1\|_2^2 = O_p\left(n^{-1} m_{\max}^2 \left(1 + n^{-1} m_{\max}^{3/2} \lambda_{\max}\right)^2\right) = O_p(n^{-1} m_{\max}^2)$$
$$\|\boldsymbol{\beta}_1^* - \tilde{\boldsymbol{\beta}}_1\|_{L_2}^2 \asymp \frac{1}{m_{\max}} \|\boldsymbol{\alpha}_1^* - \tilde{\boldsymbol{\alpha}}_1\|_2^2 = O_p\left(n^{-1} m_{\max}\right),$$

where we use Assumption A.6 and B-spline Property D.1.2. From the proof of Theorem 1 in Antoniadis et al. (2012), we have,

$$\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_{L_2} = O_p\left(n^{-1} m_{\max}^{3/2} \lambda_{\max} + \rho_n\right),$$

and (D.2.5) follows immediately. □

D.3 Proof of Theorem 5.5.1, Part 2

To prove Part 2 of Theorem 5.5.1, we can repeat the proof of Part 1 of Theorem 5.5.1 but now using Lemma D.3.1 instead of Lemma D.2.1 giving the maximal distance between Y_2^* and $\hat{Y}_2^*$ responses. The proof of Lemma D.3.1 needs two further lemmas: Lemma D.3.2 on the uniform consistency of the initial estimators $\hat{m}_1$ and $\hat{\sigma}_1$ as estimators for m and σ; and Lemma D.3.3 on the uniform consistency of $\hat{F}$ as estimator of F. The proof of Lemma D.3.2 is included, that of Lemma D.3.3 follows along the lines of a similar result (in the kernel estimation context) in Van Keilegom and Akritas (1999). The details of the proof of Lemma D.3.3 are not given but we do give and prove, in Lemma D.3.4, the key result that is needed to modify their result to our P-spline setting.

Lemma D.3.1. *If Assumptions A, B and C hold,*

$$\max_{1 \le i \le n} \mid \hat{Y}_{2i}^* - Y_{2i}^* \mid = O_p(a_n) = o_p(1),$$

where $a_n = n^{-1/2}(\log n)^{1/2} + n^{-1} m_{max}^{3/2} \lambda_{max} + \rho_n +$

$$m_{max}^{-1/2}\left(\sup_{\mathbf{u},\mathbf{x}}\left\{\tau_1(\mathbf{u},\mathbf{x}) \sup_{t \le \tau_1(\mathbf{u},\mathbf{x})} |\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + \kappa(\mathbf{u},\mathbf{x}) + \kappa_\sigma(\mathbf{u},\mathbf{x})\right\}\right).$$

Method 2 uses (5.4.4) and (3.2.3) as initial estimates for $m(\mathbf{u},\mathbf{x})$ and $\sigma^2(\mathbf{u},\mathbf{x})$. We therefore need, in the proof of Theorem 5.5.1, Part 2, the consistency results given in Lemma D.3.2.

Lemma D.3.2. *Under Assumptions A, B.1 and B.2, we have,*

$$(a) \sup_{\mathbf{u},\mathbf{x}} |\hat{m}_1(\mathbf{u},\mathbf{x}) - m(\mathbf{u},\mathbf{x})| = O_p\Bigg(n^{-1/2} + n^{-1} m_{max}^{3/2} \lambda_{max} + \rho_n + m_{max}^{-1/2} \Bigg(\sup_{\mathbf{u},\mathbf{x}} \Bigg\{ \tau_1(\mathbf{u},\mathbf{x}) \sup_{t \le \tau_1(\mathbf{u},\mathbf{x})} |\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + \kappa(\mathbf{u},\mathbf{x}) \Bigg\} \Bigg) \Bigg).$$

$$(b) \max_{1\le i\le n} | \hat{Y}^*_{1i,\sigma^2} - Y^*_{1i,\sigma^2} | = O_p\Bigg(n^{-1/2} + n^{-1} m_{max}^{3/2} \lambda_{max} + \rho_n + \sup_{\mathbf{u},\mathbf{x}} \Bigg\{ \tau_1(\mathbf{u},\mathbf{x}) \sup_{t \le \tau_1(\mathbf{u},\mathbf{x})} |\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + m_{max}^{-1/2} \kappa(\mathbf{u},\mathbf{x}) + \kappa_\sigma(\mathbf{u},\mathbf{x}) \Bigg\} \Bigg)$$

where $Y^*_{1i,\sigma^2} = \dfrac{\Delta_i (Z_i - m(\mathbf{U}_i,\mathbf{X}_i))^2}{G(Z_i|\mathbf{U}_i,\mathbf{X}_i)}$.

$$(c) \sup_{\mathbf{u},\mathbf{x}} |\hat{\sigma}_1(\mathbf{u},\mathbf{x}) - \sigma(\mathbf{u},\mathbf{x})| = O_p\Bigg(n^{-1/2} + n^{-1} m_{max}^{3/2} \lambda_{max} + \rho_n + m_{max}^{-1/2} \Big(\sup_{\mathbf{u},\mathbf{x}} \{ \tau_1(\mathbf{u},\mathbf{x}) \sup_{t \le \tau_1(\mathbf{u},\mathbf{x})} |\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + m_{max}^{-1/2} \kappa(\mathbf{u},\mathbf{x}) + \kappa_\sigma(\mathbf{u},\mathbf{x}) \} \Big) \Bigg)$$

where $Y^*_{1i,\sigma^2} = \dfrac{\Delta_i (Z_i - m(\mathbf{U}_i,\mathbf{X}_i))^2}{G(Z_i|\mathbf{U}_i,\mathbf{X}_i)}$.

Proof of Lemma D.3.2(a). Since the X_p are bounded (see Assumption A.3), we have,

$$\sup_{\mathbf{u},\mathbf{x}} |\hat{m}_1(\mathbf{u},\mathbf{x}) - m(\mathbf{u},\mathbf{x})| \lesssim \sum_{p=1}^{d} \|\hat{\beta}_{1p} - \beta_p\|_{L_\infty} \le \sum_{p=1}^{d} \|\hat{\beta}_{1p} - \tilde{\beta}_{1p}\|_{L_\infty} + \sum_{p=1}^{d} \|\tilde{\beta}_{1p} - \beta_p\|_{L_\infty}.$$

By property D.1.4, we have $\|\hat{\beta}_{1p} - \tilde{\beta}_{1p}\|_{L_\infty} \lesssim m_{\max}^{-1/2} \|\hat{\beta}_{1p} - \tilde{\beta}_{1p}\|_{L_2}$. Using the intermediate results stated in the proof of Theorem 5.5.1, part 1, we obtain that

$$\|\hat{\beta}_{1p} - \tilde{\beta}_{1p}\|_{L_\infty} = O_p\Bigg(n^{-1/2} + m_{\max}^{-1/2} \Bigg(\sup_{\mathbf{u},\mathbf{x}} \Bigg\{ \tau_1(\mathbf{u},\mathbf{x}) \sup_{t \le \tau_1(\mathbf{u},\mathbf{x})} |\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + \kappa(\mathbf{u},\mathbf{x}) \Bigg\} \Bigg) \Bigg).$$

By Lemma A.10 of Huang et al. (2004), we have,

$$\|\tilde{\boldsymbol{\beta}}_{1,reg} - \boldsymbol{\beta}\|_{L_\infty} = O_p(\rho_n),$$

where $\tilde{\beta}_{1p,reg}(u_p) = \mathbf{B}(u_p)(\mathbf{R}^T\mathbf{R})\mathbf{R}\mathbf{M}$ is the expectation of the regular spline estimator (i.e. $\lambda_1 = \ldots = \lambda_d = 0$). From the proof of Theorem 2 in Antoniadis et al. (2012), we have that

$$\tilde{\boldsymbol{\beta}}_1 = \left(1 - O_p(n^{-1}m_{\mathsf{max}}^{3/2}\lambda_{\mathsf{max}})\right)\tilde{\boldsymbol{\beta}}_{1,reg}.$$

Since each spline $\tilde{\beta}_p$ is a continuous function on the compact set $\mathcal{U}_p$, each spline $\tilde{\beta}_p$ is bounded and $\|\tilde{\boldsymbol{\beta}}_{1,reg}\|_{L_\infty} = O_P(1)$. We therefore conclude that

$$\|\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}\|_{L_\infty} = O_p(\rho_n + n^{-1}m_{\mathsf{max}}^{3/2}\lambda_{\mathsf{max}}),$$

The result of Lemma D.3.2(a) now follows.

Proof of Lemma D.3.2(b). Lemma D.3.2(b) is for $\sigma(\mathbf{u},\mathbf{x})$ what Lemma D.2.1 is for $m(\mathbf{u},\mathbf{x})$. Again we consider two cases: Z_i exceeds or does not exceed $\tau_1(\mathbf{U}_i,\mathbf{X}_i)$. Suppose first that $Z_i \le \tau_1(\mathbf{U}_i,\mathbf{X}_i)$, then we write,

$$\begin{aligned} &|\ \hat{Y}^*_{1i,\sigma^2} - Y^*_{1i,\sigma^2}\ | \\ &\le \left|\hat{m}_1^2(\mathbf{U}_i,\mathbf{X}_i) - m^2(\mathbf{U}_i,\mathbf{X}_i)\right| + 2Z_i\left|\hat{m}_1(\mathbf{U}_i,\mathbf{X}_i) - m(\mathbf{U}_i,\mathbf{X}_i)\right| \\ &\qquad + (Z_i - m(\mathbf{U}_i,\mathbf{X}_i))^2\left|\hat{G}(Z_i|\mathbf{U}_i,\mathbf{X}_i) - G(Z_i|\mathbf{U}_i,\mathbf{X}_i)\right| \end{aligned}$$

Since $\hat{m}^2(\mathbf{u},\mathbf{x}) - m^2(\mathbf{u},\mathbf{x}) = \{\hat{m}(\mathbf{u},\mathbf{x}) - m(\mathbf{u},\mathbf{x})\}\{\hat{m}(\mathbf{u},\mathbf{x}) + m(\mathbf{u},\mathbf{x})\}$, we get from the uniform convergence of $\hat{m}(\mathbf{u},\mathbf{x})$ to $m(\mathbf{u},\mathbf{x})$, that the rate of the first and second term on the right-hand side are both equal to the rate obtained in Lemma D.3.2(a). The third term on the right hand side is bounded in probability by $\sup_{t\le\tau_1(\mathbf{U}_i,\mathbf{X}_i)}|\hat{G}(t|\mathbf{U}_i,\mathbf{X}_i) - G(t|\mathbf{U}_i,\mathbf{X}_i)|$.

Next, suppose $Z_i > \tau_1(\mathbf{U}_i,\mathbf{X}_i)$, then we can write,

$$|\hat{Y}^*_{1i,\sigma^2} - Y^*_{1i,\sigma^2}| \le |\hat{Y}^*_{1i,\sigma^2} - \tilde{Y}^*_{1i,\sigma^2}| + |\tilde{Y}^*_{1i,\sigma^2} - Y^*_{1i,\sigma^2}|$$

where $\tilde{Y}^*_{1i,\sigma^2} = Y^*_{1i,\sigma^2}1_{\{Z_i\le\tau_1(\mathbf{U}_i,\mathbf{X}_i)\}} + (Z_i - m^2(\mathbf{U}_i,\mathbf{X}_i))^2 1_{\{Z_i>\tau_1(\mathbf{U}_i,\mathbf{X}_i)\}}$. Analogue to the second part of the proof of Lemma D.2.1, we use κ_σ to bound the difference between $\hat{Y}^*_{1i,\sigma^2}$ and Y^*_{1i,σ^2} in the truncation area. For the estimation of the mean of Y, the transformation formula when Z_i lies in the truncation area is Z_i, whereas in this case, the transformation formula is $(Z_i - \hat{m}_1(\mathbf{U}_i,\mathbf{X}_i))^2$ and therefore also involves an estimator $\hat{m}_1$. The variable $\tilde{Y}^*_{1i,\sigma^2}$ is introduced to make the transition from $\hat{Y}^*_{1i,\sigma^2} \equiv (Z_i - \hat{m}_1(\mathbf{U}_i,\mathbf{X}_i))^2$ via $\tilde{Y}^*_{1i,\sigma^2} \equiv (Z_i - m(\mathbf{U}_i,\mathbf{X}_i))^2$ to Y^*_{1i,σ^2}. We get

$$\mathbb{E}|\tilde{Y}^*_{1i,\sigma^2} - Y^*_{1i,\sigma^2}| \le \sup_{\mathbf{u},\mathbf{x}}\kappa_\sigma(\mathbf{u},\mathbf{x}),$$

and,

$$
\begin{aligned}
&|\hat{Y}^*_{1i,\sigma^2} - \tilde{Y}^*_{1i,\sigma^2}| \\
&\leq 2Z_i \left|\hat{m}_1(\mathbf{U}_i, \mathbf{X}_i) - m(\mathbf{U}_i, \mathbf{X}_i)\right| + \left|\hat{m}_1^2(\mathbf{U}_i, \mathbf{X}_i) - m^2(\mathbf{U}_i, \mathbf{X}_i)\right| \\
&= O_p\Bigg(n^{-1/2} + n^{-1} m_{\mathsf{max}}^{3/2} \lambda_{\mathsf{max}} + \rho_n \\
&\qquad + m_{\mathsf{max}}^{-1/2} \left(\sup_{\mathbf{u},\mathbf{x}} \left\{ \tau_1(\mathbf{u},\mathbf{x}) \sup_{t \leq \tau_1(\mathbf{u},\mathbf{x})} |\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + \kappa(\mathbf{u},\mathbf{x}) \right\} \right) \Bigg).
\end{aligned}
$$

Proof of Lemma D.3.2(c). Following the same steps as in the proof of Theorem 5.5.1, Part 1, we can, using the result of Lemma D.3.2(b), derive the L_2-distance between $\hat{\sigma}^2$ and σ^2. Analogous to Lemma D.3.2(a), the L_∞-distance then follows. Since $\hat{\sigma}_1 - \sigma = (\hat{\sigma}_1^2 - \sigma^2)/(\hat{\sigma}_1 + \sigma)$, it follows from the convergence of $\hat{\sigma}_1^2(\mathbf{u}, \mathbf{x})$ to $\sigma^2(\mathbf{u}, \mathbf{x}) > 0$, that the rate is maintained for $\hat{\sigma}_1 - \sigma$. □

Lemma D.3.3. *If assumptions A, B and C hold, then, for $t < S$, we have,*

$$
\begin{aligned}
&\hat{F}(t) - F(t) = O_p\Bigg(n^{-1/2} (\log n)^{1/2} + n^{-1} m_{max}^{3/2} \lambda_{max} + \rho_n + \\
&m_{max}^{-1/2} \left[\sup_{\mathbf{u},\mathbf{x}} \left\{ \tau_1(\mathbf{u},\mathbf{x}) \sup_{t \leq \tau_1(\mathbf{u},\mathbf{x})} |\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + \kappa(\mathbf{u},\mathbf{x}) + \kappa_\sigma(\mathbf{u},\mathbf{x}) \right\} \right] \Bigg).
\end{aligned}
$$

Lemma D.3.4. *Suppose $\beta_p \in C^r([a_p, b_p])$ for each $1 \leq p \leq d$. Then under Assumptions A and B, we have,*

$$
\begin{aligned}
&\|\hat{\boldsymbol{\beta}}_1^{(v)} - \boldsymbol{\beta}^{(v)}\|_{L_\infty} = O_p\Bigg(n^{-1/2} m_{max}^{v} + n^{-1} m_{max}^{3/2} \lambda_{max} + m_{max}^{v-r} \\
&+ m_{max}^{v-1/2} \left[\sup_{\mathbf{u},\mathbf{x}} \left\{ \tau_1(\mathbf{u},\mathbf{x}) \sup_{t \leq \tau_1(\mathbf{u},\mathbf{x})} |\hat{G}(t|\mathbf{u},\mathbf{x}) - G(t|\mathbf{u},\mathbf{x})| + \kappa(\mathbf{u},\mathbf{x}) \right\} + \rho_n \right] \Bigg),
\end{aligned}
$$

where $\boldsymbol{\beta}^{(v)} = \left(\frac{\partial^v \beta_1}{\partial u_1^v}, \ldots, \frac{\partial^v \beta_d}{\partial u_d^v} \right)^T$ and $\hat{\boldsymbol{\beta}}_1^{(v)} = \left(\frac{\partial^v \hat{\beta}_{11}}{\partial u_1^v}, \ldots, \frac{\partial^v \hat{\beta}_{1d}}{\partial u_d^v} \right)^T$ are the vectors of the v-th order derivative functions for $v = 0, \ldots, r-1$.

Proof of Lemma D.3.4. We first note that the v-th derivative of the B-spline function $\hat{\boldsymbol{\beta}}_{1p}(u_p) = \sum_{l=1}^{m_p} \hat{\alpha}_{1p,l} B_{pl}(u_p, q_p)$ of degree q_p is a B-spline function of degree $q_p - v$ given by (see De Boor (1978)),

$$
\hat{\boldsymbol{\beta}}_1^{(v)} = K_p^v \, \mathbf{b}(u_p, q - v)^T \mathbf{D}_v \hat{\boldsymbol{\alpha}}_{1p}, \tag{D.3.1}
$$

where $\mathbf{b}(u_p, q-v) = (B_{1p}(u_p, q_p - v), \ldots, B_{m_p-1,p}(u_p, q_p - v))^T$ is the vector of the $K_p + q_p - v$ B-spline basis functions of degree $q_p - v$ with knots $\boldsymbol{\xi}_p$ i.e., for $v = 1$, we have,

$$\begin{aligned}\hat{\beta}_{1p}^{(1)}(u_p) &= K_p \sum_{l=1}^{m_p-1} (\hat{\alpha}_{1p,l-1} - \hat{\alpha}_{1p,l}) B_{pl}(u_p, q_p - 1) = K_p \mathbf{b}(u_p, q-1)^T \mathbf{D}_1 \hat{\boldsymbol{\alpha}}_{1p} \\ &= K_p \left(\mathbf{b}(u_p, q-1)^T \hat{\boldsymbol{\alpha}}_{1[-1]} - \mathbf{b}(u_p, q-1)^T \hat{\boldsymbol{\alpha}}_{1[-m]} \right)\end{aligned}$$

where $\hat{\boldsymbol{\alpha}}_{1[-1]} = (\hat{\boldsymbol{\alpha}}_{12}, \ldots \hat{\boldsymbol{\alpha}}_{1m})$, $\boldsymbol{\alpha}_{1[-m]} = (\hat{\boldsymbol{\alpha}}_{11}, \ldots \hat{\boldsymbol{\alpha}}_{1,m-1})$. Representation (D.3.1) implies that the v-th derivative of β_p is again a spline function with coefficient vector $K_p \mathbf{D}_v \hat{\boldsymbol{\alpha}}_{1p}$. As a consequence we have, using Property D.1.2, that

$$\|\hat{\boldsymbol{\beta}}_1^{(v)} - \tilde{\boldsymbol{\beta}}_1^{(v)}\|_{L_2} = O_p(m_{\mathsf{max}}^{v-1/2} \|\hat{\boldsymbol{\alpha}}_1 - \tilde{\boldsymbol{\alpha}}_1\|_2). \tag{D.3.2}$$

We now use the fact that there exists a spline function (see Corollary 6.21 and (2.120) of Theorem 2.59 in Schumaker (2007)) $\zeta_p(u_p) = \sum_{l=1}^{m_p} c_{pl} B_{pl}(u_p, q_p)$ of degree q_p with equidistant knots $\boldsymbol{\xi}_p$ and coefficient vector $\mathbf{c}_p = (c_{1p}, \ldots, c_{m_p p})^T$ such that

$$\|\tilde{\boldsymbol{\beta}}_1^{(v)} - \boldsymbol{\zeta}^{(v)}\|_{L_2} = O_p(m_{\mathsf{max}}^{v} \rho_n + n^{-1} m_{\mathsf{max}}^{3/2} \lambda_{\mathsf{max}}). \tag{D.3.3}$$

To show the validity of (D.3.3), we proceed as follows. By Lemma A.7 of Huang et al. (2004), we have that $\|\tilde{\boldsymbol{\alpha}}_{1,reg} - \mathbf{c}\|_2 = O(m_{\mathsf{max}}^{1/2} \rho_n)$, using a similar argument as before we find, $\|\tilde{\boldsymbol{\beta}}_{1,reg}^{(v)} - \boldsymbol{\zeta}^{(v)}\|_{L_2} = O_p(m_{\mathsf{max}}^{v} \rho_n)$. Using the relationship

$$\tilde{\boldsymbol{\beta}}_1^{(v)} = \left(1 - O_p(n^{-1} m_{\mathsf{max}}^{3/2} \lambda_{\mathsf{max}})\right) \tilde{\boldsymbol{\beta}}_{1,reg}^{(v)}.$$

and the fact that $\boldsymbol{\beta}_{1,reg}^{(v)}$ is bounded on a compact region, we have $\|\boldsymbol{\beta}_{1,reg}^{(v)}\|_{L_2} = O_p(1)$ and (D.3.3) follows. Also note (Schumaker (2007)) that ζ_p satisfies

$$\|\beta_p^{(v)} - \zeta_p^{(v)}\|_{L_\infty} = O(m_p^{v-r}). \tag{D.3.4}$$

The rates in (D.3.2)-(D.3.4) provide the key for the proof. Indeed

$$\|\hat{\boldsymbol{\beta}}_1^{(v)} - \boldsymbol{\beta}^{(v)}\|_{L_\infty} \leq \|\hat{\boldsymbol{\beta}}_1^{(v)} - \boldsymbol{\zeta}^{(v)}\|_{L_\infty} + \|\boldsymbol{\zeta}^{(v)} - \boldsymbol{\beta}^{(v)}\|_{L_\infty}. \tag{D.3.5}$$

For the second term in (D.3.5) we use (D.3.4). For the first term, note that

$$\|\hat{\boldsymbol{\beta}}_1^{(v)} - \boldsymbol{\zeta}^{(v)}\|_{L_\infty} \lesssim m_{\mathsf{max}}^{-1/2} \|\hat{\boldsymbol{\beta}}_1^{(v)} - \boldsymbol{\zeta}^{(v)}\|_{L_2} \tag{D.3.6}$$

and that

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_1^{(v)} - \boldsymbol{\zeta}^{(v)}\|_{L_2} &\leq \|\hat{\boldsymbol{\beta}}_1^{(v)} - \tilde{\boldsymbol{\beta}}^{(v)}\|_{L_2} + \|\tilde{\boldsymbol{\beta}}_1^{(v)} - \boldsymbol{\zeta}^{(v)}\|_{L_2} \\ &= O_p(m_{\mathsf{max}}^{v-1/2} \|\hat{\boldsymbol{\alpha}}_1 - \tilde{\boldsymbol{\alpha}}_1\|_2 + m_{\mathsf{max}}^{v} \rho_n + n^{-1} m_{\mathsf{max}}^{3/2} \lambda_{\mathsf{max}}).\end{aligned} \tag{D.3.7}$$

The result now follows from the rate obtained for $\|\hat{\boldsymbol{\alpha}}_1 - \tilde{\boldsymbol{\alpha}}_1\|_2$ in Theorem 5.5.1, Part 1 in combination with (D.3.2)-(D.3.7).

□

Proof of Lemma D.3.1. We first note that $\sup_{\mathbf{u},\mathbf{x}} |\hat{m}_1(\mathbf{u},\mathbf{x}) - m(\mathbf{u},\mathbf{x})|$ and $\sup_{\mathbf{u},\mathbf{x}} |\hat{\sigma}_1(\mathbf{u},\mathbf{x}) - \sigma(\mathbf{u},\mathbf{x})|$ are both $O_p(a_n)$ by Lemma D.3.2.

We write,

$$\begin{aligned}
\hat{Y}^*_{2i} - Y^*_{2i} &= \hat{m}_1(\mathbf{U}_i,\mathbf{X}_i) - m(\mathbf{U}_i,\mathbf{X}_i) \\
&\quad + \frac{\hat{\sigma}_1(\mathbf{U}_i,\mathbf{X}_i)}{1-\hat{F}(\hat{E}_i^T)} \int_{\hat{E}_i^T}^{\hat{S}_i} sd\hat{F}(s) - \frac{\sigma(\mathbf{U}_i,\mathbf{X}_i)}{1-F(E_i^T)} \int_{E_i^T}^{S_i} sdF(s) \\
&= \{\hat{m}_1(\mathbf{U}_i,\mathbf{X}_i) - m(\mathbf{U}_i,\mathbf{X}_i)\} && \text{(D.3.8)} \\
&\quad + \frac{\hat{\sigma}_1(\mathbf{U}_i,\mathbf{X}_i) - \sigma(\mathbf{U}_i,\mathbf{X}_i)}{1-\hat{F}(\hat{E}_i^T)} \int_{\hat{E}_i^T}^{\hat{S}_i} sd\hat{F}(s) && \text{(D.3.9)} \\
&\quad + \frac{\sigma(\mathbf{U}_i,\mathbf{X}_i)\{\hat{F}(\hat{E}_i^T) - F(E_i^T)\}}{\{1-\hat{F}(\hat{E}_i^T)\}\{1-F(E_i^T)\}} \int_{\hat{E}_i^T}^{\hat{S}_i} sd\hat{F}(s) && \text{(D.3.10)} \\
&\quad + \frac{\sigma(\mathbf{U}_i,\mathbf{X}_i)}{1-F(E_i^T)} \left\{ \int_{\hat{E}_i^T}^{E_i^T} sd\hat{F}(s) + \int_{E_i^T}^{S_i} sd(\hat{F}(s) - F(s)) + \int_{S_i}^{\hat{S}_i} sd\hat{F}(s) \right\}. && \text{(D.3.11)}
\end{aligned}$$

We first consider the three integrals in (D.3.11). Using integration by part, we have,

$$\begin{aligned}
\int_{\hat{E}_i^T}^{E_i^T} sd\hat{F}(s) &= E_i^T \hat{F}(E_i^T) - \hat{E}_i^T \hat{F}(\hat{E}_i^T) - \int_{\hat{E}_i^T}^{E_i^T} \hat{F}(s)ds \\
&= E_i^T\{\hat{F}(E_i^T) - F(E_i^T)\} + \{E_i^T F(E_i^T) - \hat{E}_i^T F(E_i^T)\} + \hat{E}_i^T\{F(E_i^T) - \hat{F}(\hat{E}_i^T)\} \\
&\qquad - \int_{\hat{E}_i^T}^{E_i^T} \hat{F}(s)ds. && \text{(D.3.12)}
\end{aligned}$$

For the first term of (D.3.12), using Lemma D.3.3, we conclude that

$$\left| E_i^T\{\hat{F}(E_i^T) - F(E_i^T)\} \right| = |E_i^T| O_p(a_n) = O_p(a_n).$$

Since $|E_i^T| \leq \{\sigma(\mathbf{U}_i,\mathbf{X}_i)\}^{-1}\{|\min(Z_i, \tau_2(\mathbf{U}_i,\mathbf{X}_i))| + |m(\mathbf{U}_i,\mathbf{X}_i)|\} < \infty$. To get a consistency rate for the second and the fourth term of (D.3.12), note that

$$\begin{aligned}
&\hat{E}_i^T - E_i^T \\
&= \frac{\min(\tau_2(\mathbf{U}_i,\mathbf{X}_i),Z_i) - \hat{m}_1(\mathbf{U}_i,\mathbf{X}_i)}{\hat{\sigma}_1(\mathbf{U}_i,\mathbf{X}_i)} - \frac{\min(\tau_2(\mathbf{U}_i,\mathbf{X}_i),Z_i) - m(\mathbf{U}_i,\mathbf{X}_i)}{\sigma(\mathbf{U}_i,\mathbf{X}_i)} \\
&= \frac{1}{\sigma(\mathbf{U}_i,\mathbf{X}_i)\hat{\sigma}(\mathbf{U}_i,\mathbf{X}_i)}\Big[\min(\tau_2(\mathbf{U}_i,\mathbf{X}_i),Z_i)\big\{\sigma(\mathbf{U}_i,\mathbf{X}_i) - \hat{\sigma}_1(\mathbf{U}_i,\mathbf{X}_i)\big\} \\
&\qquad - \sigma(\mathbf{U}_i,\mathbf{X}_i)\big\{\hat{m}_1(\mathbf{U}_i,\mathbf{X}_i) - m(\mathbf{U}_i,\mathbf{X}_i))\big\} \\
&\qquad + m(\mathbf{U}_i,\mathbf{X}_i)\big\{\hat{\sigma}_1(\mathbf{U}_i,\mathbf{X}_i) - \sigma(\mathbf{U}_i,\mathbf{X}_i)\big\}\Big].
\end{aligned}$$

It then follows from Lemma D.3.2 and the convergence of $\hat{\sigma}_1(\mathbf{u},\mathbf{x})$ to $\sigma(\mathbf{u},\mathbf{x}) > 0$ that

$$|\hat{E}_i^T - E_i^T| = O_p(a_n),$$

which gives the rate for the second and the fourth term of (D.3.12). For the third term of (D.3.12), we have that

$$\hat{F}(\hat{E}_i^T) - F(E_i^T) = \{\hat{F}(\hat{E}_i^T) - F(\hat{E}_i^T)\} + \{F(\hat{E}_i^T) - F(E_i^T)\}.$$

Lemma D.3.3 can be used for the first summand. For the second summand, we use a first order Taylor approximation and write,

$$\begin{aligned}
F(\hat{E}_i^T) - F(E_i^T) = &\Bigg(-\frac{\hat{m}_1(\mathbf{U}_i,\mathbf{X}_i) - m(\mathbf{U}_i,\mathbf{X}_i)}{\hat{\sigma}_1(\mathbf{U}_i,\mathbf{X}_i)} \\
&- \frac{\{\hat{\sigma}_1(\mathbf{U}_i,\mathbf{X}_i) - \sigma(\mathbf{U}_i,\mathbf{X}_i)\}\{\min(\tau_2(\mathbf{U}_i,\mathbf{X}_i),Z_i) - m(\mathbf{U}_i,\mathbf{X}_i)\}}{\hat{\sigma}_1(\mathbf{U}_i,\mathbf{X}_i)\sigma_1(\mathbf{U}_i,\mathbf{X}_i)}\Bigg) f_\varepsilon(\theta),
\end{aligned}$$

with f_ε the density of ε and for some θ between $\frac{\min(\tau_2(\mathbf{U}_i,\mathbf{X}_i),Z_i) - \hat{m}_1(\mathbf{U}_i,\mathbf{X}_i)}{\hat{\sigma}_1(\mathbf{U}_i,\mathbf{X}_i)}$ and $\frac{\min(\tau_2(\mathbf{U}_i,\mathbf{X}_i),Z_i) - m(\mathbf{U}_i,\mathbf{X}_i)}{\sigma(\mathbf{U}_i,\mathbf{X}_i)}$. By the convergence of $\hat{\sigma}_1(\mathbf{u},\mathbf{x})$ to $\sigma(\mathbf{u},\mathbf{x}) > 0$ and the fact that $\sup_e |ef_\varepsilon(e)| < \infty$, we get

$$F(\hat{E}_i^T) - F(E_i^T) = O_p(a_n). \tag{D.3.13}$$

We conclude that

$$\Big|\hat{E}_i^T\{F(E_i^T) - \hat{F}(\hat{E}_i^T)\}\Big| = O_p(a_n),$$

where we use that by Lemma D.3.2, $|\hat{E}_i^T| = |E_i^T| + O_p(a_n) < \infty$. Based on the analysis of (D.3.12) we conclude for the first term of (D.3.11),

$$\frac{\sigma(\mathbf{U}_i,\mathbf{X}_i)}{1 - F(E_i^T)}\int_{\hat{E}_i^T}^{E_i^T} s d\hat{F}(s) = O_p(a_n). \tag{D.3.14}$$

In a similar way, we obtain for the third term of (D.3.11)

$$\frac{\sigma(\mathbf{U}_i, \mathbf{X}_i)}{1 - F(E_i^T)} \int_{\hat{S}_i^T}^{S_i^T} s d\hat{F}(s) = O_p(a_n). \quad \text{(D.3.15)}$$

For the second integral in (D.3.11), we use partial integration and Lemma D.3.3 to obtain

$$\begin{aligned} \int_{E_i^T}^{S_i^T} sd(\hat{F}(s) - F(s)) = \quad & S_i^T\{\hat{F}(S_i^T) - F(S_i^T)\} - E_i^T\{\hat{F}(E_i^T) - F(E_i^T)\} \\ & - \int_{E_i^T}^{S_i^T} \{\hat{F}(s) - F(s)\} ds = O_p(a_n). \end{aligned}$$

The terms (D.3.8)-(D.3.10) are more easy to handle. For (D.3.8) we use Lemma D.3.2(a). For (D.3.9) and (D.3.10) we need that

$$\int_{\hat{E}_i^T}^{\hat{S}_i} s d\hat{F}(s) = O_p(1). \quad \text{(D.3.16)}$$

To show (D.3.16), note that, using similar reasoning as in Heuchenne and Van Keilegom (2007), we can prove that

$$\int_{E_i^T}^{S_i} s d\hat{F}(s) = O_p(1).$$

Combining this result with the rates obtained in (D.3.14) and (D.3.15) yields,

$$\int_{\hat{E}_i^T}^{\hat{S}_i} s d\hat{F}(s) = O_p(1).$$

By the convergence of $\hat{F}(\hat{E}_i^T)$ to $F(E_i^T) < 1$ (D.3.13), we get that (D.3.9) and (D.3.10) are both $O_p(a_n)$. □

D.4 Proof of Theorem 5.5.2

Proof of Theorem 5.5.2. We prove the asymptotic normality of the P-spline estimator $\hat{\boldsymbol{\beta}}_1$ for method 1 by proving that for $1 \leq p \leq d$,

$$\left\{s.e.\left(\beta_{jp}^*(u_p) \mid \mathcal{X}_n\right)\right\}^{-1}\left\{\beta_{jp}^*(u_p) - \tilde{\beta}_{jp}(u_p)\right\} \xrightarrow{d} \mathrm{N}(0,1) \quad \text{(D.4.1)}$$

$$\left\{s.e.\left(\beta_{jp}^*(u_p) \mid \mathcal{X}_n\right)\right\}^{-1}\Big\{(\hat{\beta}_{1p}(u_p) - \beta_{1p}^*(u_p)) + (\tilde{\beta}_{1p}(u_p) - \beta_p(u_p))\Big\} \xrightarrow{p} 0. \quad \text{(D.4.2)}$$

The proof of (D.4.1) is based on the proof given in Antoniadis et al. (2012) where some steps can be simplified due to the independence of the observations.

Let $\mathbf{B}_p(\mathbf{u})$ be the column vector representing the p-th row of $\mathbf{B}(\mathbf{u})$.

$$\mathbf{B}_p^T(\mathbf{u})(\boldsymbol{\alpha}^* - \tilde{\boldsymbol{\alpha}}) = \sum_{i=1}^n \mathbf{B}_p^T(\mathbf{u})(\mathbf{R}^T\mathbf{R} + \mathbf{Q}_{\boldsymbol{\lambda}})^{-1}\mathbf{R}_i(Y_{1i}^* - M_{1i}) = \sum_{i=1}^n d_i\xi_i,$$

where $d_i^2 = \sigma_{1,i}^2\{\mathbf{B}_p^T(\mathbf{u})(\mathbf{R}^T\mathbf{R} + \mathbf{Q}_{\boldsymbol{\lambda}})^{-1}\mathbf{R}_i\}^2$ and $\xi_i = \sigma_{1,i}^{-2}(Y_{1i}^* - M_{1i})$. Conditioning on $\mathcal{X}_n$ the ξ_i are independent with mean 0 and variance 1. To prove the asymptotic normality of the P-spline estimator we verify that the Lindeberg condition,

$$\frac{\max d_i^2}{\sum_{i=1}^n d_i^2} \xrightarrow{p} 0,$$

is satisfied, then,

$$\frac{\sum_{i=1}^n d_i\xi_i}{\sqrt{\sum_{i=1}^n d_i^2}} \xrightarrow{d} \mathsf{N}(0,1).$$

For any $\boldsymbol{\omega} = (\boldsymbol{\omega}_0^T, \ldots, \boldsymbol{\omega}_d^T)^T$ with $\boldsymbol{\omega}_p = (\omega_{p1}, \ldots, \omega_{pm_p})^T$, and especially for $\boldsymbol{\omega} = \{\mathbf{R}^T\mathbf{R} + \mathbf{Q}_{\boldsymbol{\lambda}})^{-1}\mathbf{B}_p(\mathbf{u})\}$, we have by the Cauchy-Schwarz inequality,

$$\begin{aligned}\boldsymbol{\omega}^T\mathbf{R}_i\mathbf{R}_i^T\boldsymbol{\omega} &= \left\{\sum_{p=0}^d X_{ip}\sum_{l=1}^{m_p}\omega_{pl}B_{pl}(U_{ip};q_p)\right\}^2 \\ &\le \left(\sum_{p=0}^d X_{ip}^2\right)\left[\sum_{p=0}^d\left\{\sum_{l=1}^{m_p}\omega_{pl}B_{pl}(U_{ip};q_p)\right\}^2\right].\end{aligned}$$

Set $g_{\boldsymbol{\omega},p}(u;q_p) = \sum_{l=1}^{m_p}\omega_{pl}B_{pl}(u_p;q_p)$ for $p = 0,\ldots,d$. By Assumption (B3) and Properties D.1.2 and D.1.4,

$$\boldsymbol{\omega}^T\mathbf{R}_i\mathbf{R}_i^T\boldsymbol{\omega} \lesssim \sum_{p=0}^d \|g_{\boldsymbol{\omega},p}\|_\infty^2 \lesssim m_{\mathsf{max}}\sum_{p=0}^d \|g_{\boldsymbol{\omega},p}\|_{L_2}^2 \asymp \|\boldsymbol{\omega}\|_2^2. \tag{D.4.3}$$

From Lemmas A.1 and A.2 in Huang et al. (2004), we know that except on an event with probability tending to zero, $n^{-1}\sum_{i=1}^n(\sum_{p=0}^d X_{ip}g_{\boldsymbol{\omega},p}(U_{ip};q_p))^2 \asymp m_{\mathsf{max}}^{-1}\|\boldsymbol{\omega}\|_2^2$. Thus,

$$\begin{aligned}\boldsymbol{\omega}^T\sum_{i=1}^n\left\{\mathbf{R}_i\mathbf{R}_i^T\sigma_{1,i}^2\right\}\boldsymbol{\omega} &\ge n\min_{1\le i\le n}\sigma_{1,i}^2 n^{-1}\sum_{i=1}^n\Big(\sum_{p=0}^d X_{ip}g_{\boldsymbol{\omega},p}(U_{ip};q_p)\Big)^2 \\ &\gtrsim m_{\mathsf{max}}^{-1}n\|\boldsymbol{\omega}\|_2^2. \end{aligned}\tag{D.4.4}$$

Combining (D.4.3) and (D.4.4), we find that except on an event whose probability tends to zero, we have,

$$\frac{\max_i(\sigma_{1,i}^2\omega^T\mathbf{R}_i\mathbf{R}_i^T\omega)}{\boldsymbol{\omega}^T(\sum_{i=1}^n\sigma_{1,i}^2\mathbf{R}_i\mathbf{R}_i^T)\boldsymbol{\omega}} \lesssim n^{-1}m_{\mathsf{max}}.$$

By Assumption (A6), it follows that the Lindeberg assumption is fulfilled and hence the normality result in (D.4.1) follows.

We continue with the proof of (D.4.2). Since we assume that $\sigma^2_{1,i}$ is bounded away from zero and ∞, we have,

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{\beta}^*_{1p}(\mathbf{u}) \mid \mathcal{X}_n) &= \mathrm{Cov}\left(\mathbf{B}_p^T(\mathbf{u})\boldsymbol{\alpha}^* \mid \mathcal{X}_n\right) \\
&= \mathbf{B}(\mathbf{u})\big(\mathbf{R}^T\mathbf{R}+\mathbf{Q}_{\boldsymbol{\lambda}}\big)^{-1}\Big(\sum_{i=1}^n \mathbf{R}_i\mathbf{R}_i^T\sigma^2_{1,i}\Big)\big(\mathbf{R}^T\mathbf{R}+\mathbf{Q}_{\boldsymbol{\lambda}}\big)^{-1}\mathbf{B}_p(\mathbf{u}) \\
&\gtrsim \mathbf{B}_p^T(\mathbf{u})\big(\mathbf{R}^T\mathbf{R}+\mathbf{Q}_{\boldsymbol{\lambda}}\big)^{-1}\mathbf{R}^T\mathbf{R}\big(\mathbf{R}^T\mathbf{R}+\mathbf{Q}_{\boldsymbol{\lambda}}\big)^{-1}\mathbf{B}_p(\mathbf{u}) \\
&\asymp \frac{n}{m_{\max}}\mathbf{B}_p^T(\mathbf{u})\big(\mathbf{R}^T\mathbf{R}+\mathbf{Q}_{\boldsymbol{\lambda}}\big)^{-1}\big(\mathbf{R}^T\mathbf{R}+\mathbf{Q}_{\boldsymbol{\lambda}}\big)^{-1}\mathbf{B}_p(\mathbf{u}) \\
&\gtrsim \frac{n}{m_{\max}}\left(\frac{1}{\lambda_{\max}(\mathbf{R}^T\mathbf{R}+\mathbf{Q}_{\boldsymbol{\lambda}})}\right)^2\sum_{l=1}^{m_p}B^2_{pl}(\mathbf{u}) \\
&\gtrsim \frac{n}{m_{\max}}\left(\frac{1}{\frac{n}{m_{\max}}\left(1+\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\right)}\right)^2\frac{1}{m_p} \\
&\asymp \frac{1}{n}\left(1+\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\right)^{-2},
\end{aligned}$$

where we use the Cauchy-Schwarz inequality,

$$1=\left(\sum_{l=1}^{m_p}B_{pl}(\mathbf{u})\right)^2 \le \sum_{l=1}^{m_p}B^2_{pl}(\mathbf{u})\sum_{l=1}^{m_p}1 = m_p\sum_{l=1}^{m_p}B^2_{pl}(\mathbf{u}),$$

and the upper bound for the largest eigenvalue $\lambda_{\max}(\mathbf{R}^T\mathbf{R}+\mathbf{Q}_{\boldsymbol{\lambda}})$:

$$\begin{aligned}
\lambda_{\max}(\mathbf{R}^T\mathbf{R}+\mathbf{Q}_{\boldsymbol{\lambda}}) &= \|\mathbf{R}^T\mathbf{R}+\mathbf{Q}_{\boldsymbol{\lambda}}\|_2 \le \|\mathbf{R}^T\mathbf{R}\|_2 + \|\mathbf{Q}_{\boldsymbol{\lambda}}\|_2 \\
&\lesssim \frac{n}{m_{\max}} + \sqrt{\sum_{p=1}^d \|\mathbf{Q}_{\boldsymbol{\lambda}}\|_\infty} \lesssim \frac{n}{m_{\max}} + \sqrt{d}\lambda_{\max}m_{\max}^{1/2}\max_{1\le p\le d}4^{k_p} \\
&\lesssim \frac{n}{m_{\max}}\left(1+\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\right).
\end{aligned}$$

By Property D.1.4 of B-splines and Assumption (A5),

$$\begin{aligned}
\hat{\beta}_{1p}(u_p)-\beta^*_{1p}(u_p) &\le \sup_{u\in\mathcal{U}}|\hat{\beta}_{1p}(u_p)-\beta^*_{1p}(u_p)| = \|\hat{\beta}_{1p}-\beta^*_{1p}\|_\infty \\
&\lesssim \left(\frac{1}{m_p}\right)^{1/2}\|\hat{\beta}_{1p}-\beta^*_{1p}\|_{L_2} \asymp \left(\frac{1}{m_{\max}}\right)^{1/2}\|\hat{\beta}_{1p}-\beta^*_{1p}\|_{L_2}.
\end{aligned}$$

We conclude,

$$\frac{\hat{\beta}_{1p}(u_p)-\beta^*_{1p}(u_p)}{s.e.\left(\beta^*_{1p}(u_p)\mid\mathcal{X}_n\right)} \lesssim \left(\frac{n}{m_{\max}}\right)^{1/2}\left(1+\frac{m_{\max}^{3/2}\lambda_{\max}}{n}\right)\|\hat{\beta}_{1p}-\beta^*_{1p}\|_{L_2},$$

and

$$\frac{\tilde{\beta}_{1p}(u_p) - \beta_p(u_p)}{s.e.\left(\beta^*_{1p}(u_p) \mid \mathcal{X}_n\right)} \lesssim n^{1/2}\left(1 + \frac{m_{\mathsf{max}}^{3/2}\lambda_{\mathsf{max}}}{n}\right)\|\tilde{\beta}_{1p} - \beta_p\|_{L_\infty}.$$

From assumption D.1 it follows that these two terms converge to zero as n goes to ∞. The proof for method 2 is exactly the same but we do not look at the difference $\tilde{\beta}_{2p} - \beta_p$. □

Bibliography

Abrevaya, J. (1999). Rank regression for current-status data: asymptotic normality. *Stat. Probab. Lett.*, 43(3):275–287.

Abrevaya, J. and Huang, J. (2005). On the bootstrap of the maximum score estimator. *Econometrica*, 73(4):1175–1204.

Antoniadis, A., Gijbels, I., and Verhasselt, A. (2012). Variable selection in varying-coefficient models using P-splines. *J. Comput. Graph. Stat.*, 21(3):638–661.

Antoniadis, A., Grégoire, G., and McKeague, I. W. (2004). Bayesian estimation in single index models. *Stat. Sin.*, 14:1147–1164.

Aragón, J. and Quiroz, A. J. (1995). Rank regression for current status data. *Stat. Probab. Lett.*, 24(3):251–256.

Ayer, M., Brunk, H., Ewing, G., Reid, W., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Stat.*, 26:641–647.

Balabdaoui, F., Durot, C., and Jankowski, H. (2016). Least squares estimation in the monotone single index model. *arXiv preprint arXiv:1610.06026*.

Balabdaoui, F., Groeneboom, P., and Hendrickx, K. (2018). Score estimation in the monotone single index model. Submitted for publication.

Banerjee, M., Durot, C., and Sen, B. (2017). Divide and conquer in non-standard problems and the super-efficiency phenomenon. Submitted.

Banerjee, M. and Wellner, J. (2001). Likelihood ratio tests for monotone functions. *Ann. Stat.*, 29:1699–1731.

Banerjee, M. and Wellner, J. A. (2005). Confidence intervals for current status data. *Scand. J. Stat.*, 32(3):405–424.

Barlow, R., Bartholomew, D., Bremner, J., and Brunk, H. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression*. John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical report, University of California, Berkeley*.

Bravo, F. (2014). Varying coefficients partially linear models with randomly censored data. *Ann. I. Stat. Math.*, 66(2):383–412.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, 66:429–436.

Cai, J., Fan, J., Zhou, H., and Zhou, Y. (2007). Hazard models with varying coefficients for multivariate failure time data. *Ann. Stat.*, 35(1):324–354.

Cai, Z., Fan, J., and Li, R. (1999). Generalized varying-coefficient models. *Department of Statistics, UCLA*.

Cai, Z., Fan, J., and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *J. Am. Stat. Assoc.*, 95(451):941–956.

Caplehorn, J. and Bell, J. (1991). Methadone dosage and retention of patients in maintenance treatment. *Med J Aust*, 154(3):195–199.

Cavanagh, C. and Sherman, R. P. (1998). Rank estimators for monotonic index models. *J. Econometrics*, 84(2):351–381.

Chambers, J., Cleveland, W., Kleiner, B., and Tukey, P. (1983). *Graphical Methods for Data Analysis*. Duxbury Press, Boston.

Cosslett, S. (1987). Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models. *Econometrica*, 55(3):559–585.

Cosslett, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica*, 51(3):765–782.

Cosslett, S. R. (2007). Efficient estimation of semiparametric models by smoothed maximum likelihood. *Int. Econ. Rev.*, 48(4):1245–1272.

Cox, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc. Ser. B*, 34:187–220. With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.

Cui, X., Härdle, W. K., and Zhu, L. (2011). The efm approach for single-index models. *Ann. Statist.*, 39(3):1658–1688.

De Boor, C. (1978). A practical guide to splines. *Springer, New York*.

Delecroix, M., Härdle, W., and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *J. Multivariate Anal.*, 86(2):213–226.

Ding, Y. and Nan, B. (2011). A sieve M-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *Ann. Stat.*, 39(6):3032–3061.

Durot, C., Groeneboom, P., and Lopuhaä, H. P. (2013). Testing equality of functions under monotonicity constraints. *J. Nonparametr. Stat.*, 25(4):939–970.

Durot, C. and Reboul, L. (2010). Goodness-of-fit test for monotone functions. *Scand. J. Stat.*, 37(3):422–441.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Stat.*, 7:1–26.

Eilers, P. H. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–102.

El Ghouch, A. and Van Keilegom, I. (2008). Nonparametric regression with dependent censored data. *Scand. J. Stat.*, 35:228–247.

Eubank, R. L. and Speckman, P. (1990). Curve fitting by polynomial-trigonometric regression. *Biometrika*, 77(1):1–9.

Fahrmeir, L. and Tutz, G. (2001). Multivariate statistical modelling based on generalized linear models, second edition. *Springer*.

Fan, J. and Gijbels, I. (1994). Censored regression: local linear approximations and their applications. *J. Am. Stat. Assoc.*, 89(426):560–570.

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Stat. Interface*, 1(1):179–195.

Finkelstein, D. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 42(4):845–854.

Finkelstein, D. M. and Wolfe, R. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, 41(4):933–945.

Giné, E. and Nickl, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Univ. Press, Cambridge.

Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, 39:125–153 (1957).

Groeneboom, P. (1996). Lectures on inverse problems. In *Lectures on probability theory and statistics (Saint-Flour, 1994)*, volume 1648 of *Lecture Notes in Math.*, pages 67–164. Springer, Berlin.

Groeneboom, P. (2012). Likelihood ratio type two-sample tests for current status data. *Scand. J. Stat.*, 39:645–662.

Groeneboom, P. and Hendrickx, K. (2017a). Current status linear regression. *Accepted for publication in Ann. Stat.*

Groeneboom, P. and Hendrickx, K. (2017b). The nonparametric bootstrap for the current status model. *Electron. J. Stat.*, 11(2):3446–3484.

Groeneboom, P. and Hendrickx, K. (2018a). Confidence intervals for the current status model. *Scand. J. Stat.*, 45(1):135–163. doi: 10.1111/sjos.12294.

Groeneboom, P. and Hendrickx, K. (2018b). Estimation in monotone single index models. *Stat. Neerl.* doi:10.1111/stan.12138.

Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation under Shape Constraints*. Cambridge Univ. Press, Cambridge.

Groeneboom, P., Jongbloed, G., and Witte, B. I. (2010). Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model. *Ann. Stat.*, 38(1):352–387.

Groeneboom, P. and Ketelaars, T. (2011). Estimators for the interval censoring problem. *Electron. J. Stat.*, 5:1797–1845.

Groeneboom, P. and Wellner, J. (1992). *Information bounds and nonparametric maximum likelihood estimation*, volume 19 of *DMV Seminar*. Birkhäuser Verlag, Basel.

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multivariate Anal.*, 32:177–203.

Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Ann. Stat.*, 20:675–694.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *J. Econometrics*, 35(2-3):303–316.

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Stat.*, 21(1):157–178.

Hastie, T. and Tibshirani, R. (1993). Varying coefficient models. *J. R. Stat. Soc. Ser. B*, 55:757–796.

Hendrickx, K., Janssen, P., and Verhasselt, A. (2017). Penalized spline estimation in varying coefficient models with censored data. *TEST*. doi:10.1007/s11749-017-0574-y.

Heuchenne, C. and Van Keilegom, I. (2007). Polynomial regression with censored data based on preliminary nonparametric estimation. *Ann. I. Stat. Math.*, 59(2):273–297.

Hooke, R. and Jeeves, T. A. (1961). "direct search"solution of numerical and statistical problems. *Journal of the ACM*, 8(2):212–229.

Hoover, D., Rice, J., Wu, C., and Yang, L. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822.

Hristache, M., Juditsky, A., and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Stat.*, 29(3):595–623.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Stat.*, 24(2):540–568.

Huang, J. and Wellner, J. (1993). Regression models with interval censoring. *Proceedings of the Kolmogorov Seminar, Euler Mathematics Institute. St. Petersburg, Russia*.

Huang, J. Z., Wu, C. O., and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Stat. Sin.*, 14(3):763–788.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *J. Econometrics*, 58(1-2):71–120.

Jin, Z., Lin, Y., and Ying, Z. (2006). On least-squares regression with censored data. *Biometrika*, 93(1):147–161.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, 53:457–481.

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *J. R. Stat. Soc. Ser. A*, 154(3):371–412. With discussion.

Keiding, N., Begtrup, K., Scheike, T., and Hasibeder, G. (1996). Estimation from current status data in continuous time. *Lifetime Data Anal.*, 2:119–129.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Stat.*, 18:191–219.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, 61(2):387–421.

Kosorok, M. (2008). Bootstrapping the Grenander estimator. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, volume 1 of *Inst. Math. Stat. Collect.*, pages 282–292. Inst. Math. Stat., Beachwood, OH.

Koul, H., Susarla, V., and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *Ann. Stat.*, 9(6):1276–1288.

Kuchibhotla, A. K. and Patra, R. K. (2017). Efficient estimation in single index models through smoothing splines. available at `https://arxiv.org/abs/1612.00068`.

Lagakos, S., Barraj, L. M., and Degruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to aids. *Biometrika*, 75(3):515–523.

Lee, Y., Mammen, E., and Park, B. (2012). Flexible generalized varying coefficient regression models. *Ann. Stat.*, 40(3):1906–1933.

Leurgans, S. (1987). Linear models, random censoring and synthetic data. *Biometrika*, 74(2):301–309.

Li, G. and Zhang, C.-H. (1998). Linear regression with interval censored data. *Ann. Stat.*, 26(4):1306–1327.

Ma, Y. and Wei, Y. (2012). Analysis on censored quantile residual life model via spline smoothing. *Stat. Sin.*, 22(1):47.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *J. Econometrics*, 3(3):205–228.

Meeker, W. (1987). Limited failure population life tests: application to integrated circuit reliability. *Technometrics*, 29(1):51–65.

Murphy, S. A., van der Vaart, A. W., and Wellner, J. A. (1999). Current status regression. *Math. Methods Stat.*, 8(3):407–425.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142.

Nahman, N. S., Middendorf, D. F., Bay, W. H., McElligott, R., Powell, S., and Anderson, J. (1992). Modification of the percutaneous approach to peritoneal dialysis catheter placement under peritoneoscopic visualization: clinical results in 78 patients. *J Am. Soc. Nephrol.*, 3(1):103–107.

Newey, W. K. and Stoker, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica*, 61(5):1199–1223.

Pollard, D. (1984). *Convergence of stochastic processes*. Springer Series in Statistics. Springer-Verlag, New York.

Prakasa Rao, B. (1969). Estimation of a unimodal density. *Sankhyā Ser. A*, 31:23–36.

Rabinowitz, D., Tsiatis, A., and Aragon, J. (1995). Regression with interval-censored data. *Biometrika*, 82(3):501–513.

Robertson, T., Wright, F., and Dykstra, R. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.

Schumaker, L. (2007). Spline functions: Basic theory, 3th edition. *Cambridge University Press, New York*.

Sen, B., Banerjee, M., and Woodroofe, M. (2010). Inconsistency of bootstrap: the Grenander estimator. *Ann. Stat.*, 38:1953–1977.

Sen, B. and Xu, G. (2015). Model based bootstrap methods for interval censored data. *Comput. Stat. Data An.*, 81:121–129.

Shen, S.-L., Cui, J.-L., and Wang, C.-W. (2016). Re-weighting estimation of the coefficients in the varying coefficient model with heteroscedastic errors. *J. Stat. Comput. Simul.*, 86(10):1921–1935.

Shen, X. (2000). Linear regression with current status data. *J. Am. Stat. Assoc.*, 95(451):842–852.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica*, 61(1):123–137.

Stone, C. (1992). Optimal global rates of convergence for nonparametric regression. *Ann. Stat.*, 10(4):1040–1053.

Subbotin, V. (2007). Asymptotic and bootstrap properties of rank regressions. Available at SSRN: `https://ssrn.com/abstract=1028548`.

Tanaka, H. (2008). Semiparametric least squares estimation of monotone single index models and its application to the iterative least squares estimation of binary choice models. Technical report.

Torczon, V. (1997). On the convergence of pattern search algorithms. *SIAM J. Optim.*, 7(1):1–25.

van de Geer, S. (2000). *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, Cambridge.

van der Vaart, A. (1991). On differentiable functionals. *Ann. Stat.*, 19(1):178–204.

van der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

van Eeden, C. (1956). Maximum likelihood estimation of ordered probabilities. *Nederl. Akad. Wetensch. Proc. Ser. A.*, 18:444–455.

Van Keilegom, I. and Akritas, M. G. (1999). Transfer of tail information in censored regression models. *Ann. Stat.*, 27(5):1745–1784.

Wang, S., Nan, B., Zhu, J., and Beer, D. G. (2008). Doubly penalized Buckley-James method for survival data with high-dimensional covariates. *Biometrics*, 64(1):132–140, 323.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Wei, L. J. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Stat. Med.*, 11 14-15:1871–9.

Wu, C. O., Chiang, C., and Hoover, D. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Am. Stat. Assoc.*, 93(444):1388–1402.

Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *J. Multivariate Anal.*, 97(5):1162–1184.

Yang, S., El Ghouch, A., and Van Keilegom, I. (2014). Varying coefficient models having different smoothing variables with randomly censored data. *Electron. J. Stat.*, 8:226–252.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *J. Am. Stat. Assoc.*, 97(460):1042–1054.

Zheng, Z. (1987). A class of estimators of the parameters in linear regression with censored data. *Acta mathematicae applicatae sinica*, 3(3):231–241.

Zheng, Z. (2008). A class of estimators of the mean survival time from interval censored data with application to linear regression. *Applied Mathematics-A Journal of Chinese Universities*, 23(4):377–390.

Zhou, M. (1991). Some properties of the kaplan-meier estimator for independent non-identically distributed random variables. *Ann. Stat.*, 19(4):2266–2274.