Received: 6 July 2017 | Revised: 9 January 2018 | Accepted: 9 March 2018

DOI: 10.1111/stan.12138

ORIGINAL ARTICLE

WILEY

Estimation in monotone single-index models

Piet Groeneboom[1] | Kim Hendrickx[2]

[1]Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands

[2]I-BioStat, Hasselt University, Hasselt, Belgium

Correspondence
Piet Groeneboom, Delft Institute of Applied Mathematics, Delft University of Technology, Delft, The Netherlands.
Email: P.Groeneboom@tudelft.nl

Funding information
Research Foundation Flanders (FWO), Grant/Award Number: 11W7315N; IAP Research Network P7/06 of the Belgian State (Belgian Science Policy); VSC-Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government-department EWI

Single-index models are popular regression models that are more flexible than linear models and still maintain more structure than purely nonparametric models. We consider the problem of estimating the regression parameters under a monotonicity constraint on the unknown link function. In contrast to the standard approach of using smoothing techniques, we review different "non-smooth" estimators that avoid the difficult smoothing parameter selection. For about 30 years, one has had the conjecture that the profile least squares estimator is an $\sqrt{n}$-consistent estimator of the regression parameter, but the only non-smooth argmin/argmax estimators that are actually known to achieve this $\sqrt{n}$-rate are not based on the nonparametric least squares estimator of the link function. However, solving a score equation corresponding to the least squares approach results in $\sqrt{n}$-consistent estimators. We illustrate the good behavior of the score approach via simulations. The connection with the binary choice and current status linear regression models is also discussed.

KEYWORDS
least squares, monotone link function, single-index model

1 | THE SINGLE-INDEX MODEL

Suppose that Y is a response variable and that $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is a d-dimensional covariate ($d \geq 1$). The semiparametric single-index model is given by

$$Y = \psi_0\left(\boldsymbol{\alpha}_0^T \boldsymbol{X}\right) + \varepsilon, \tag{1}$$

where $\varepsilon \sim F_0$ is a random error term with $\mathbb{E}(\varepsilon|\boldsymbol{X}) = 0$ almost surely, ψ_0 is an unknown function, and $\boldsymbol{\alpha}_0$ is an unknown regression parameter in $\mathbb{R}^d$. These models are more flexible than standard linear regression models and have, on the other hand, more structure than completely nonparametric models.

In this paper, we impose a monotonicity constraint on the link function ψ_0. Shape-constrained inferences have gained a lot of popularity in recent years. Several statistical applications are based on imposing constraints that occur from the problem under study and need algorithms for shape-constrained regression or density estimation. Monotonicity, convexity, or concavity constraints arise naturally with consumption or production functions, growth curves, and dose response models. A summary of nonparametric estimation under shape constraints can be found in Groeneboom and Jongbloed (2014).

Model (1), when ψ_0 is an unknown monotone function, is also known as the monotone single-index model, and many econometric models as well as censored regression models and various duration models fit into this framework. We next describe how the binary choice model and the current status linear regression model are special cases of the monotone single-index model 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

$$Y = \begin{cases} 1, & \text{if } \boldsymbol{\alpha}_0^T\boldsymbol{X} \geq \varepsilon \\ 0, & \text{else,} \end{cases} \tag{2}$$

where $\boldsymbol{\alpha}_0^T\boldsymbol{X}$ presents the utility score and ε, assumed to be independent of $\boldsymbol{X}$, is the disturbance term. 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 single-index model (1) with ψ_0 equal to the (unknown) distribution function F_0 of ε, since

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

The current status linear regression model. A frequently encountered problem in regression analysis is that a variable of interest (which can be thought of as an event time) is not observed directly but only known to lie before or after some random censoring time. This type of censored data is known as current status data and arises commonly in reliability and survival studies especially when testing is destructive. More formally, instead of observing the vector $(Y, \boldsymbol{X})$ as in (1), a vector $(T, \Delta, \boldsymbol{X})$ is observed, where T is a censoring variable and $\Delta = 1_{Y \leq T}$. Since

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

the current status linear regression model is a special case of the single-index model (1) with response $\tilde{Y} = \Delta$, covariate vector $\tilde{\boldsymbol{X}} = (T, \boldsymbol{X})^T$, and $\tilde{\boldsymbol{\alpha}}_0 = (1, \boldsymbol{\alpha}_0^T)$.

The models are prototypes of semiparametric models, where one has a nonparametric component given by the "link function" ψ_0, which is a distribution function (or 1 minus a distribution function) in the current status regression model, and a parametric part. The difficulty of the model is that the parametric part is "inside" the nonparametric part; one has to bypass the nonparametric function ψ_0, which cannot be estimated at rate $\sqrt{n}$, to get to the parametric part.

This is very different for the well-known Cox proportional hazards model for current status data. In this case, the log likelihood is of the form

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

where Λ is the baseline cumulative hazard function. Now, $\boldsymbol{\alpha}_0$ does not appear in the argument of a function F, which is not $\sqrt{n}$ estimable, and we can estimate Λ and $\boldsymbol{\alpha}_0$ separately. In this case, it was shown in Huang (1996) that one can use the nonparametric maximum likelihood estimator of Λ and then use profile likelihood to estimate $\boldsymbol{\alpha}_0$ efficiently at rate $\sqrt{n}$. However, for the ordinary current status regression model, it is still unknown whether a similar estimation method gives an $\sqrt{n}$-consistent estimate of $\boldsymbol{\alpha}_0$.

1.1 | Identifiability

Identification 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 $(\boldsymbol{\alpha}_0, \psi_0)$, however, cannot be estimated in the single-index model. 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

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

Even if the distribution of $(Y, \boldsymbol{X})$ is known, the above model cannot be distinguished from model (1) unless restrictions on 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

$$\left\{\boldsymbol{\alpha} \in \mathbb{R}^d : \alpha_1 = 1\right\} \quad \text{or} \quad \left\{\boldsymbol{\alpha} \in \mathbb{R}^d : \|\boldsymbol{\alpha}_0\| = 1, \alpha_{01} \geq 0\right\},$$

where $\|\cdot\|$ denotes the Euclidean norm and α_{01} is the first component of $\boldsymbol{\alpha}_0$, ensures scale identification of the model. The first parametrization is used in Sherman (1993), whereas examples of the second parametrization are found in Härdle, Hall, and Ichimura (1993) and Hristache, Juditsky, and Spokoiny (2001), among others. Note that in the special case of the current status linear regression model, the first component of the covariate vector corresponds to the censoring variable T with a coefficient equal to 1, and therefore, the current status regression model is identified without further restrictions on the parameter space.

1.2 | Efficient information

The log likelihood for the model of one observation is given by

$$\ell_{\boldsymbol{\alpha},\psi}(\boldsymbol{x}, y) \stackrel{\text{def}}{=} \log\left\{f_{\varepsilon|\boldsymbol{X}}\left(y - \psi(\boldsymbol{\alpha}^T\boldsymbol{x})\right) f_{\boldsymbol{X}}(\boldsymbol{x})\right\},$$

where $f_{\varepsilon|\boldsymbol{X}}$ is the conditional density of ε given $\boldsymbol{X} = \boldsymbol{x}$ and $f_{\boldsymbol{X}}$ is the density of $\boldsymbol{X}$.

The partial derivative w.r.t. $\boldsymbol{\alpha}$ of $\ell_{\boldsymbol{\alpha},\psi}$ is given by

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

Let $\{\psi_\eta : \eta \in (-1, 1)\}$ be a path in the collection $\{\psi : \psi \text{ is increasing}\}$, 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

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

To obtain the efficient score function, we must solve

$$\mathbb{E}\left[\left\{\frac{\boldsymbol{X}\psi'(\boldsymbol{\alpha}^T\boldsymbol{X})f'_{\epsilon|\boldsymbol{X}}\left(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right)}{f_{\epsilon|\boldsymbol{X}}\left(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right)}\right.\right.$$
$$\left.\left.-\frac{a_*(\boldsymbol{\alpha}^T\boldsymbol{X})f'_{\epsilon|\boldsymbol{X}}\left(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right)}{f_{\epsilon|\boldsymbol{X}}\left(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right)}\right\}\frac{a(\boldsymbol{\alpha}^T\boldsymbol{X})f'_{\epsilon|\boldsymbol{X}}\left(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right)}{f_{\epsilon|\boldsymbol{X}}\left(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right)}\right]=0, \tag{3}$$

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

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

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

$$\tilde{\ell}_{\alpha,\psi}(\boldsymbol{x},y)=\frac{y-\psi(\boldsymbol{\alpha}^T\boldsymbol{x})}{\sigma^2(\boldsymbol{x})}\psi'(\boldsymbol{\alpha}^T\boldsymbol{x})\left\{\boldsymbol{x}-\frac{\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})\boldsymbol{X}|\boldsymbol{\alpha}^T\boldsymbol{X}=\boldsymbol{\alpha}^T\boldsymbol{x}\right\}}{\mathbb{E}\left\{\sigma^{-2}(\boldsymbol{X})|\boldsymbol{\alpha}^T\boldsymbol{X}=\boldsymbol{\alpha}^T\boldsymbol{x}\right\}}\right\}, \tag{4}$$

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

$$\mathbb{E}\left[\frac{\left\{\boldsymbol{X}\psi'(\boldsymbol{\alpha}^T\boldsymbol{X})-a_*(\boldsymbol{\alpha}^T\boldsymbol{X})\right\}f'_{\epsilon|\boldsymbol{X}}\left(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right)^2}{f_{\epsilon|\boldsymbol{X}}(Y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X}))^2}a(\boldsymbol{\alpha}^T\boldsymbol{X})\right]$$
$$=\mathbb{E}\left[\left\{\boldsymbol{X}\psi'(\boldsymbol{\alpha}^T\boldsymbol{X})-a_*(\boldsymbol{\alpha}^T\boldsymbol{X})\right\}\frac{\left\{y-\psi(\boldsymbol{\alpha}^T\boldsymbol{X})\right\}^2}{\sigma^4(\boldsymbol{X})}a(\boldsymbol{\alpha}^T\boldsymbol{X})\right]$$
$$=\mathbb{E}\left[\mathbb{E}\left\{\frac{\left\{\boldsymbol{X}\psi'(\boldsymbol{\alpha}^T\boldsymbol{X})-a_*(\boldsymbol{\alpha}^T\boldsymbol{X})\right\}}{\sigma^2(\boldsymbol{X})}\middle|\boldsymbol{\alpha}^T\boldsymbol{X}\right\}a(\boldsymbol{\alpha}^T\boldsymbol{X})\right].$$

Note that

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

This means that (3) is solved for

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

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

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

Efficiency calculations for the binary choice model are given in Cosslett (1987), and the efficient score function $\tilde{\ell}_{\boldsymbol{\alpha}_0,F_0}$ is equal to

$$\tilde{\ell}_{\boldsymbol{\alpha}_0,F_0}(y,x) = \left\{x - \mathbb{E}(X|\boldsymbol{\alpha}^T\boldsymbol{X} = \boldsymbol{\alpha}^T\boldsymbol{x})\right\} f(\boldsymbol{\alpha}^T\boldsymbol{x}) \left\{ \frac{y}{F(\boldsymbol{\alpha}^T\boldsymbol{x})} - \frac{1-y}{1-F(\boldsymbol{\alpha}^T\boldsymbol{x})} \right\}, \tag{5}$$

where $f = F'$. For the current status model, the efficient score resembles the efficient score for the binary choice model except that Y and $\boldsymbol{\alpha}^T\boldsymbol{x}$ are replaced by Δ and $t - \boldsymbol{\alpha}^T\boldsymbol{x}$, respectively (see, e.g., Huang & Wellner, 1993 or Murphy, van der Vaart, & Wellner, 1999).

The $\sqrt{n}$-consistent estimators with an asymptotic normal distribution and n times the limiting variance equal to the inverse of

$$\mathbb{E}\left(\tilde{\ell}_{\boldsymbol{\alpha},\psi}(\boldsymbol{X},Y)\tilde{\ell}_{\boldsymbol{\alpha},\psi}(\boldsymbol{X},Y)^T\right)$$

are called efficient estimators of $\boldsymbol{\alpha}_0$. In the single-index model without the monotonicity constraint on ψ_0, the efficient estimators of $\boldsymbol{\alpha}_0$ have been constructed in Ichimura (1993) and Delecroix et al. (2003). Klein and Spady (1993) developed an efficient quasi-maximum likelihood estimator for the binary choice model. An efficient estimate for the current status linear regression model based on a penalized maximum likelihood procedure is proposed in Murphy et al. (1999). Inspired by this penalized estimate, an efficient penalized least squares estimate (PLSE) is constructed for the single-index model with convex link function ψ_0 in Kuchibhotla and Patra (2017).

1.3 | Single-index regression parameter estimators

Several 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 for 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 $\sum_{i=1}^n (Y_i - \psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X}_i))^2$, when ψ_0 is estimated using a kernel estimator (depending on some bandwidth h). Härdle et al. (1993) extended the SLSE by minimizing the 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, Grégoire, and McKeague (2004), among others. The average derivative estimator proposed by Hristache et al. (2001) results in a direct estimation of the regression parameter $\boldsymbol{\alpha}_0$ and, therefore, avoids solving a hard optimization problem for 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). Smoothing techniques are needed to allow for an efficient estimation of $\boldsymbol{\alpha}_0$ in the single-index model.

Examples of M-estimators that are not based on an estimate of ψ_0 are Manski's maximum rank estimator (MRE) (Manski, 1975) for the binary choice model, the maximum rank correlation estimator (MRCE) proposed by Han (1987), and the rank estimators proposed by Cavanagh and Sherman (1998) for a more general generalized regression model under monotonicity constraints. The convergence rate for Manski's estimator is in contrast to the other estimators discussed in this section, somewhat disappointingly $n^{1/3}$ instead of the usual $\sqrt{n}$-rate.

1.4 | Aim of the paper

In this paper, we discuss the behavior of regression parameter estimators in the monotone single-index model. All the estimators are obtained using tuning-parameter-free algorithms and are derived from non-smooth and non-convex criterion functions. In Section 2, we distinguish between two different classes of estimators.

(a) Estimators with an unknown limiting distribution that depend on the behavior of the piecewise constant, monotone least squares estimator (LSE) of the link function.

(b) Estimators with a known limiting distribution that converge at the parametric rate to the true regression parameters.

Within the first class of estimators, we first discuss the profiled LSE of the regression parameter in Section 2.1. It is proved in Balabdaoui, Durot, and Jankowski (2016) that the LSE converges at least at the cube-root n rate, but its limiting distribution is still an open problem. Inspired by the rank estimator proposed in Aragón and Quiroz (1995) for the current status model, we also propose a new estimator in this class in Section 2.5.

In Section 2.2, we explain how the $\sqrt{n}$-consistent estimators proposed in Balabdaoui, Groeneboom, and Hendrickx (2017) for the monotone single-index model can be derived from a score approach. For the second class of estimators, we also describe the rank estimators proposed by Han (1987) and Cavanagh and Sherman (1998) in Section 2.3 and Section 2.4, respectively.

The remainder of Section 2 is devoted to the following.

- The asymptotic properties of the estimators. In particular, we describe a general approach in Section 2.6 for proving the asymptotic normal distribution of the estimators of the second class (b), and we discuss, in Section 2.7, the difficulties that arise when one wants to derive the limiting distribution of the estimators of the first class (a).
- The computation of the estimators.

The quality of the estimators is illustrated via simulations and a real data example in Section 3 and Section 4, respectively. 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. A slight loss of efficiency is an acceptable price to pay for a tuning-free parameter procedure that is computationally more attractive than efficient procedures (since no smoothing parameter selection is needed). Efficient estimators are moreover often based on smoothness conditions that are stronger than the conditions needed when smoothing techniques are avoided.

2 | ESTIMATORS OBTAINED WITHOUT SMOOTHING TECHNIQUES

In this section, we describe different estimators for $\boldsymbol{\alpha}_0$ based on a random sample $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ of size n from $(\boldsymbol{X}, Y)$, where $\mathbb{E}(Y|\boldsymbol{X}) = \psi_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})$ and ψ_0 belongs to the class $\mathcal{M}$ of monotone functions on $\mathbb{R}$. To ensure identifiability of the single-index model, we assume that $\boldsymbol{\alpha}_0$ is a vector of regression parameters belonging to the $(d-1)$-dimensional sphere $\mathcal{S} := \{\boldsymbol{\alpha} \in \mathbb{R}^d : \|\boldsymbol{\alpha}\| = 1\}$.

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

$$Y = \exp\left(X_1/\sqrt{2} + X_2/\sqrt{2}\right) + \varepsilon, \qquad X_1, X_2 \sim U[-1,1] \text{ and } \varepsilon \sim N(0,1). \tag{6}$$

For each of the estimators, we include figures for the criterion function as a function of θ, where θ is defined by $(\alpha_1, \alpha_2) = (\cos(\theta), \sin(\theta))$. In model (6), the true parameter value is $\theta_0 = \pi/4$.

2.1 | The least squares estimator

Consider the sum of squared errors

$$S_n(\boldsymbol{\alpha}, \psi) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \psi\left(\boldsymbol{\alpha}^T\boldsymbol{X}_i\right)\right\}^2, \tag{7}$$

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

$$(\hat{\boldsymbol{\alpha}}_n, \hat{\psi}_n) := \arg\min_{\boldsymbol{\alpha}\in S, \psi\in\mathcal{M}} S_n(\boldsymbol{\alpha}, \psi). \tag{8}$$

The LSE can be obtained as follows. For a fixed $\boldsymbol{\alpha} \in S_{d-1}$, 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 $Z_1^{\boldsymbol{\alpha}} < \cdots < Z_m^{\boldsymbol{\alpha}}$ be the corresponding ordered values. For $i = 1, \ldots, m$, 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 the 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

$$\left\{(0,0), \left(\sum_{j=1}^{i} n_j^{\boldsymbol{\alpha}}, \sum_{j=1}^{i} n_j^{\boldsymbol{\alpha}} Y_j^{\boldsymbol{\alpha}}\right), i = 1, \ldots, m\right\}. \tag{9}$$

See, for example, theorem 1.1 in Barlow, Bartholomew, Bremner, and Brunk (1972) or theorem 1.2.1 in Robertson, Wright, and Dykstra (1988). By strict convexity of $\psi \mapsto S_n(\psi, \boldsymbol{\alpha})$, the minimizer is unique at the distinct projections. We define the nonparametric least squares estimator (LSE(ψ)) $\hat{\psi}_{n\boldsymbol{\alpha}}$ by the monotone function that takes the values of this minimizer at the distinct projections and is extended to a right-continuous step function outside the set of those projections. Figure 1 shows a picture of the LSE of $\hat{\psi}_{n\boldsymbol{\alpha}_0}$ for the model described in (6). The LSE $(\hat{\boldsymbol{\alpha}}_n, \hat{\psi}_n) := (\hat{\boldsymbol{\alpha}}_n, \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n})$ is next obtained by maximizing the map $\boldsymbol{\alpha} \mapsto S_n(\boldsymbol{\alpha}, \hat{\psi}_{n\boldsymbol{\alpha}})$ over all $\boldsymbol{\alpha} \in S$.

Since the cumulative sum diagram in (9) only changes when the ordering in $\boldsymbol{\alpha}^T\boldsymbol{X}_i$ changes, the vector $(\hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_1), \ldots, \hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T\boldsymbol{X}_n))^T$ will be the same for all $\boldsymbol{\alpha}$ for which the ranks of the $\boldsymbol{\alpha}^T\boldsymbol{X}_i$ are the same. As a consequence, the criterion function $\boldsymbol{\alpha} \mapsto S_n(\boldsymbol{\alpha}, \hat{\psi}_{n\boldsymbol{\alpha}})$ is piecewise constant, and the LSE $(\hat{\boldsymbol{\alpha}}_n, \hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n})$ is not unique.

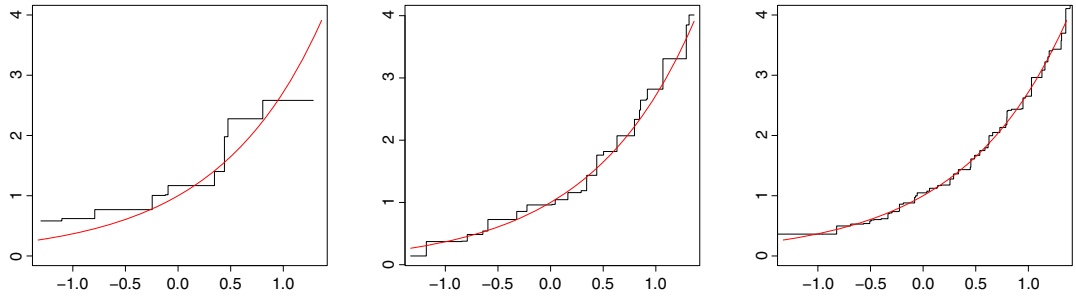

FIGURE 1 The LSE $\hat{\psi}_{n,\alpha_0}$ (black, step-wise) and the true function $\psi_0(x) = \exp(x)$ (red, solid) in model (6) for a sample of sizes (a) $n = 100$, (b) $n = 1000$, and (c) $n = 10000$

2.1.1 | The maximum likelihood estimator for the binary choice model and the current status linear regression model

Instead of considering the LSE for the binary choice model or the current status model, we consider maximizing the log likelihood of the data. In particular, for the binary choice model, the MLE $(\hat{\boldsymbol{\alpha}}_n, \hat{F}_n)$ is defined by

$$(\hat{\boldsymbol{\alpha}}_n, \hat{F}_n) := \arg \max_{\boldsymbol{\alpha} \in S, F \in \mathcal{F}} L_n(\boldsymbol{\alpha}, F),$$

where $\mathcal{F}$ is the set of all distribution functions on $\mathbb{R}$ and

$$L_n(\boldsymbol{\alpha}, F) := \frac{1}{n} \sum_{i=1}^{n} \left[Y_i \log F\left(\boldsymbol{\alpha}^T \boldsymbol{X}_i\right) + (1 - Y_i) \log \left\{ 1 - F\left(\boldsymbol{\alpha}^T \boldsymbol{X}_i\right) \right\} \right]. \tag{10}$$

Using profiled log likelihood, we first obtain the minimizer of the map $F \mapsto L_n(\boldsymbol{\alpha}, F)$ over $\mathcal{F}$, which is, in fact, the same function as the minimizer of $S_n(\boldsymbol{\alpha}, F)$. For the current status model, we replace Y_i and $\boldsymbol{\alpha}^T \boldsymbol{X}_i$ in the expression for L_n given in (10) by Δ_i and $T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i$ and maximize over $\mathbb{R} \times \mathcal{F}$ instead of $S \times \mathcal{F}$ due to the identifiability of the current status regression model.

Note that in a homoscedastic model with normal error terms, the MLE for the single-index model (1) is equivalent to the LSE. The asymptotic properties of the LSE in (8) for the monotone single-index model and the MLE in (10) for the binary choice model and the current status model are discussed in Balabdaoui et al. (2016), Cosslett (1983), and Murphy et al. (1999), respectively. All estimators converge at rate $n^{1/3}$, but their limiting distribution is still an open problem.

2.2 | The simple score estimator

Balabdaoui et al. (2017) developed a Z-estimator based on the derivative of the sum of squares $S_n(\boldsymbol{\alpha}, \hat{\psi}_{n\boldsymbol{\alpha}})$, ignoring the non-differentiability of the LSE $\hat{\psi}_{\boldsymbol{\alpha}_n}$. They first consider a local parametrization $\mathbb{S}$ mapping a subset of $\mathbb{R}^{d-1}$ to the sphere S. Examples of such parametrizations are the map

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

or the spherical coordinate system

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

The simple score estimator (SSE) is next obtained by a zero-crossing (in $\boldsymbol{\beta}$) of

$$Z_n(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{J}_{\mathbb{S}}(\boldsymbol{\beta}))^T\boldsymbol{X}_i\left\{Y_i - \hat{\psi}_{n\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}^T\boldsymbol{X}_i\right)\right\}, \tag{11}$$

where $\boldsymbol{\alpha} = \mathbb{S}(\boldsymbol{\beta})$ and $J_{\mathbb{S}} \in \mathbb{R}^{d\times d-1}$ is the matrix of the partial derivatives of $\mathbb{S} : \mathbb{R}^{d-1} \mapsto \mathcal{S} : \boldsymbol{\beta} \mapsto \mathbb{S}(\boldsymbol{\beta}) = \boldsymbol{\alpha}$. Since the LSE $\hat{\psi}_{n\boldsymbol{\alpha}}$ is the same for different $\boldsymbol{\alpha}$ values, the score criterion (11) will have discontinuities and an exact root of (11) does not always exist. The estimator is therefore defined by $\hat{\boldsymbol{\alpha}}_n = \mathbb{S}(\hat{\boldsymbol{\beta}}_n)$, where $\hat{\boldsymbol{\beta}}_n$ is a point in $\mathbb{R}^{d-1}$ such that each component of Z_n crosses through zero at $\hat{\boldsymbol{\beta}}_n$.

The SSE is, in fact, a generalization of the related estimator for the current status linear regression model proposed in Groeneboom and Hendrickx (2017) and defined by a zero-crossing of

$$Z_n'(\boldsymbol{\alpha}) := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\left\{\Delta_i - \hat{F}_{n\boldsymbol{\alpha}}\left(T_i - \boldsymbol{\alpha}^T\boldsymbol{X}_i\right)\right\}, \tag{12}$$

where $\hat{F}_{n\boldsymbol{\alpha}}$ is the MLE for F_0 when $\boldsymbol{\alpha}$ is fixed.

2.3 | The maximum rank correlation estimator

Han's MRCE 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

$$H_n(\boldsymbol{\alpha}) := \frac{1}{n(n-1)}\sum_{i\neq j}\{Y_i > Y_j\}\left\{\boldsymbol{\alpha}^T\boldsymbol{X}_i > \boldsymbol{\alpha}^T\boldsymbol{X}_j\right\}. \tag{13}$$

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

2.4 | The maximum rank estimator

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

$$R_n(\boldsymbol{\alpha}) := \frac{1}{n(n-1)}\sum_{i\neq j}M(Y_i)\left\{\boldsymbol{\alpha}^T\boldsymbol{X}_i > \boldsymbol{\alpha}^T\boldsymbol{X}_j\right\}, \tag{14}$$

where M denotes an increasing function on $\mathbb{R}$. In this paper, we investigate the behavior of the estimator when M is equal to the identity function, that is, $M(y)_{=}y$, and refer to this estimator as the MRE. Since the responses in the binary choice model and the current status model are binary, it can be shown (see Appendix, Section A.1) that the MRCE and the MRE are equivalent in these models. 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.

2.5 | The maximum rank estimator using the least squares estimator of $\boldsymbol{\psi_0}$

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

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

TABLE 1 Asymptotic variance. Monotone single-index model

Method	$\boldsymbol{\Sigma}$
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	$\mathbb{E}\left[\psi_0'(\boldsymbol{\alpha}_0^T\boldsymbol{X})\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}\left\{\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}\vert\boldsymbol{\alpha}_0^T\boldsymbol{X})\right\}^T g_0(\boldsymbol{\alpha}_0^T\boldsymbol{X})\right]$

Note. SSE = simple score estimator; MRCE = maximum rank correlation estimator; MRE = maximum rank estimator.

where $\hat{F}_{n\boldsymbol{\alpha}}$ is the MLE (see (10)) for fixed $\boldsymbol{\alpha}$. This motivates us to investigate the behavior of the estimator for the monotone single-index model, referred to as the LS-MRE, defined by the maximizer of

$$A_n(\boldsymbol{\alpha}) := \frac{1}{n(n-1)}\sum_{i\neq j}\hat{\psi}_{n\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}^T\boldsymbol{X}_i\right)\left\{\boldsymbol{\alpha}^T\boldsymbol{X}_i > \boldsymbol{\alpha}^T\boldsymbol{X}_j\right\}, \tag{15}$$

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 for ψ_0, it can be expected that similar theoretical issues appear when deriving the limiting behavior for both estimators.

2.6 | Asymptotic behavior

It has been shown in Balabdaoui et al. (2017) 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 distribution with an asymptotic variance that is larger than the efficient variance. As pointed out in a footnote on p. 361 of Cavanagh and Sherman, the expression for the asymptotic variance of the MRCE given in theorem 4 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.

Although no proof for the MRCE and the MRE has been published, we can prove that

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)\rightarrow_d N(\boldsymbol{0},\boldsymbol{V}^-\Sigma,\boldsymbol{V}^-), \tag{16}$$

where $\boldsymbol{V}^-$ is the Moore–Penrose inverse of $\boldsymbol{V}$; a sketch of the proof of (16) is given in the Appendix, Section A.2. The reason that we 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 therefore degenerate. This is also clear from its covariance matrix $\boldsymbol{V}^-\boldsymbol{\Sigma}\boldsymbol{V}^-$, which is a matrix of rank $d-1$. The expressions for $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ are summarized in Table 1 for the monotone single-index model and in Table 2 for the current status linear regression model.

The limiting distributions of the LSE and the LS-MRE are still unknown. Figures 2 and 3 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 Figures 4 and 5.

TABLE 2 Asymptotic variance. Current status linear regression model

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

Note. SSE = simple score estimator; MRCE = maximum rank correlation estimator; MRE = maximum rank estimator.

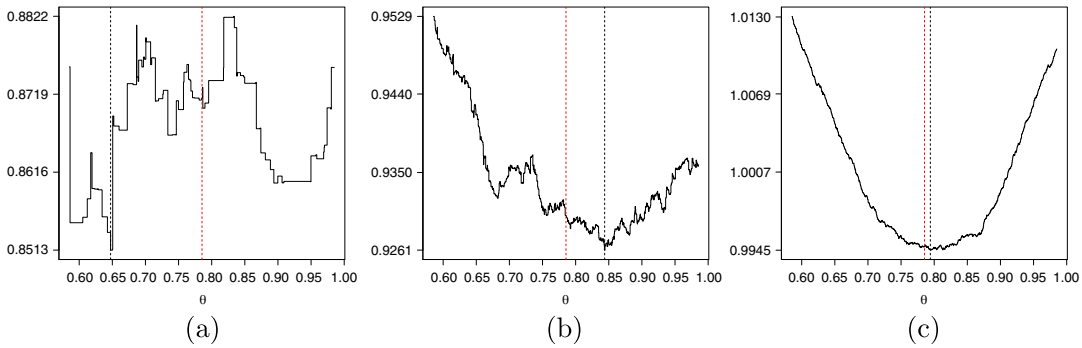

FIGURE 2 The map $\theta \mapsto S_n((\cos(\theta), \sin(\theta))^T, \hat{\psi}_{n,\boldsymbol{\alpha}})$ (black, solid) in model (6) for a sample of sizes (a) $n = 100$, (b) $n = 1000$, and (c) $n = 10000$. The vertical reference lines indicate the position of the minimizer (black, dotted) and the true parameter value $\theta_0 = \pi/4$ (red, dotted)

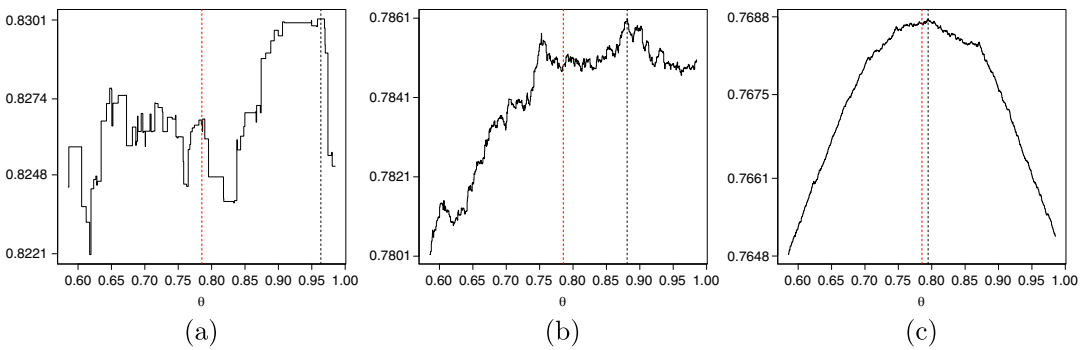

FIGURE 3 The map $\theta \mapsto A_n((\cos(\theta), \sin(\theta)^T))$ (black, solid) in model (6) for a sample of sizes (a) $n = 100$, (b) $n = 1000$, and (c) $n = 10000$. The vertical reference lines indicate the position of the maximizer (black, dotted) and the true parameter value $\theta_0 = \pi/4$ (red, dotted)

2.7 | Difficulties with the LSE and the LS-MRE

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 SLSE proposed in Ichimura (1993), where the criterion function is given by S_n defined in (7) but with $\hat{\psi}_{n\boldsymbol{\alpha}}$ replaced by a kernel estimate

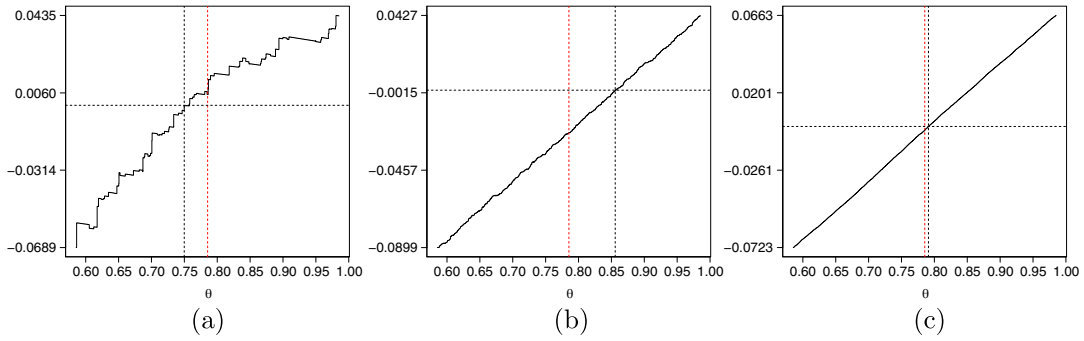

FIGURE 4 The map $\theta \mapsto Z_n((\cos(\theta), \sin(\theta)^T))$ (black, solid) in model (6) for a sample of sizes (a) $n = 100$, (b) $n = 1000$, and (c) $n = 10000$. The vertical reference lines indicate the position of the zero-crossing (black, dotted) and the true parameter value $\theta_0 = \pi/4$ (red, dotted)

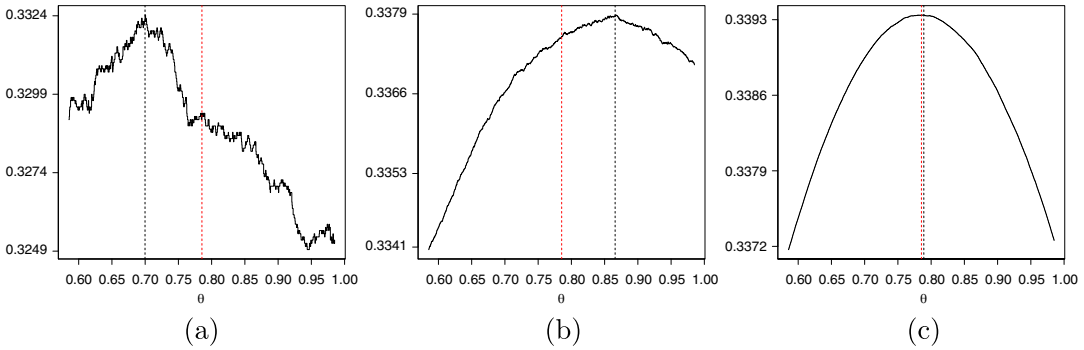

FIGURE 5 The map $\theta \mapsto H_n((\cos(\theta), \sin(\theta)^T))$ (black, solid) in model (6) for a sample of sizes (a) $n = 100$, (b) $n = 1000$, and (c) $n = 10000$. The vertical reference lines indicate the position of the maximizer (black, dotted) and the true parameter value $\theta_0 = \pi/4$ (red, dotted)

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.

As discussed in Groeneboom and Hendrickx (2017), the "canonical" approach to proofs that argmax or argmin estimates of $\boldsymbol{\alpha}_0$ are $\sqrt{n}$-consistent has been provided by Sherman (1993) for the models, considered in his paper. His theorem 1 says that $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\| = O_p(n^{-1/2})$, where $\|\cdot\|$ denotes the Euclidean norm, if $\hat{\boldsymbol{\alpha}}_n$ is the maximizer of a criterion function $\Gamma_n(\boldsymbol{\alpha})$, with population equivalent $\Gamma(\boldsymbol{\alpha})$ and

(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 $\boldsymbol{\alpha}_0$,

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

Moreover, assuming $\boldsymbol{\alpha}_0 = \mathbf{0}$ and $\Gamma(\boldsymbol{\alpha}_0) = \mathbf{0}$, he continues by making the assumption that

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

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

If we try to apply this to our situation, we first have to deal with the fact that we have to use a parametrization of the type introduced in Section 2.2 in order to consider full neighborhoods of $\boldsymbol{\beta}_0$ and nondegenerate matrices in $\mathbb{R}^{d-1}$ in the single-index model. But then, for the profile LSE estimator $\hat{\boldsymbol{\beta}}_n$, it is not clear that an expansion as given in Sherman (1993) will hold. We seem to get inevitably an extra term of order $O_p(n^{-2/3})$ in (b) (with $\boldsymbol{\alpha}_n$ and $\boldsymbol{\alpha}_0$ replaced by $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_0$), which does not fit into this framework.

On the other hand, in the expansion of our score function (11), we get that this function is in first order the sum of a term of the form

$$\phi'(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0),$$

where ϕ' is the matrix, representing the total derivative of the population equivalent score function, and a term $\boldsymbol{W}_n$ of order $O_p(n^{-1/2})$, which gives

$$\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \sim -\phi'(\boldsymbol{\beta}_0)^{-1}\boldsymbol{W}_n = O_p(n^{-1/2}),$$

and, here, extra terms of order $O_p(n^{-2/3})$ do not hurt.

Li and Zhang (1998), in their paper on smooth U-statistics estimators for the regression parameter in the current status linear regression model, conjecture in their remark 2.4 that the profile MLE (which is similar to the LSE) will be $\sqrt{n}$-convergent, but not efficient. We could not follow their argument for this conjecture. Aragón and Quiroz (1995) have a similar conjecture for their second estimate, based on the nonparametric MLE, but again, the argument for this conjecture in their comment (i) in section 4 is not at all convincing. In fact, our simulations show that the simple argmin or argmax estimates seem to be always inferior to the simple score estimates, and the pictures of the argmin-type estimates as a function of the parameter show that the behavior is a lot more irregular than the behavior of the score estimates. Hence, even if the argmin-type or argmax-type estimators would be $\sqrt{n}$-convergent, the score estimators have a better behavior.

2.8 | Computation

None of the discontinuous criterion functions in Sections 2.1–2.5 are convex. 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.

For the computation of the *M*-estimators (LSE, MRCE, MRE, and LS-MRE), we wrote C++ programs, using the pattern search algorithm proposed by Hooke and Jeeves (1961). The latter optimization method searches for the minimum of a loss function without requiring a gradient. Based on an initial starting value, $2d$ local searches of a certain step size are made in each direction to see if a lower loss function is obtained. The first move is made in the direction of the previous move. If no function decrease is found, the step size is reduced. The procedure is iterated until a convergence criterion is satisfied. The convergence of this pattern search algorithm is discussed in Torczon (1997). The root of the set of $d - 1$ score functions in Section 2.2 coincides with the minimizer of the sum of squared component score functions so that this minimization approach

can also be used to obtain the SSE. We recognize the importance of good starting values and improvements of the current approach are worth studying in future computational research.

3 | SIMULATIONS

To evaluate the finite sample behavior of the different estimators introduced in Section 2, we simulate $N = 5000$ data sets from the model

$$Y = \psi_0\left(\boldsymbol{\alpha}_0^T \boldsymbol{X}\right) + \varepsilon, \tag{17}$$

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 $i = 1, 2, 3$.

Tables 3 and 4 show the mean and n times the covariance matrix of the estimates for sample sizes $n = 100, 500, 1000, 5000$, and 100000 for the uniform, respectively normal, simulation setting. For the SSE, the MRCE, and the MRE, we calculated the asymptotic variances given in Table 1 to which n times the covariance matrix should converge. We however note that only the uniform model satisfies the assumptions needed to prove (16). The last column in Tables 3 and 4 contains the distance between n times the covariance matrix of the estimates and the matrix $\boldsymbol{V}^-\boldsymbol{\Sigma}\boldsymbol{V}^-$ obtained by summing up 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 6.

For both simulation settings, the results show the 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 the MRE. 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 the MRE are substantially larger than the ones for the SSE.

For the LS-MRE, n times the covariance matrices increase with increasing sample size, suggesting a slower convergence rate than the parametric $\sqrt{n}$-rate for this estimator. Table 3 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 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 estimates in the uniform model with the PLSE proposed by Kuchibhotla and Patra (2017) and the EFM estimate proposed by Cui, Härdle, and Zhu (2011). The function `simestgcv` available in the R-packages `simest` was used to obtain the PLSE.

The algorithm described on p. 1670 of Cui et al. (2011) was implemented by us in a C++ program, with a fixed tuning parameter M for the EDF approach. The computation time is considerably longer than the time required for the methods discussed in Section 2 (this observation also holds for the PLSE estimate proposed in Kuchibhotla & Patra, 2017). Moreover, the fixed-point algorithm, used in the second step of the algorithm of Cui et al. (2011), has an oscillating behavior and will certainly not converge in any monotone way; one has to wait whether it will enter at a certain point of the iterations into a sufficiently small ball around a fixed point. The version of the algorithm, kindly sent to us by Dr. Cui, depends crucially on the R algorithm `smooth.lf` in the R package `locfit`, and it is not so easy to see how the first step of the algorithm is handled by this package. We used our own C++ implementation of the algorithm, but noticed that the R implementation had a similar behavior. The R implementation of Cui et al. (2011) also uses a fixed tuning parameter M for the EDF approach. In both implementations, one has to discard non-converging runs.

TABLE 3 Simulation model ($X_i \sim U[0,1], d = 3$): 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 of the estimates and the asymptotic covariance matrix 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 for different sample sizes n with $N = 5000$. The line, preceded by ∞, gives 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	–
	1000	0.576605	0.575703	0.576451	1.276503	1.271816	1.249995	−0.645524	−0.628822	−0.618845	–
	5000	0.577477	0.576993	0.576840	1.421290	1.446315	1.409902	−0.730242	−0.691148	−0.715894	–
	10000	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
	1000	0.574333	0.576839	0.579007	0.695695	0.753258	0.700203	−0.368076	−0.322984	−0.381043	2.976074
	5000	0.576343	0.577253	0.578089	0.688215	0.709718	0.707268	−0.344059	−0.341827	−0.366363	2.914229
	10000	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
	1000	0.576239	0.576586	0.576801	0.926655	0.931801	0.938936	−0.458336	−0.465244	−0.472174	4.499525
	5000	0.577250	0.577172	0.577165	0.882133	0.881460	0.911412	−0.426668	−0.454751	−0.455531	4.270952
	10000	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
	1000	0.576398	0.576643	0.576752	0.865576	0.864699	0.874510	−0.425514	−0.435668	−0.438601	3.978506
	5000	0.577365	0.577191	0.577069	0.814612	0.820094	0.825936	−0.404390	−0.409542	−0.415915	3.729037
	10000	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	–
	1000	0.576197	0.575715	0.576921	1.241740	1.235253	1.235171	−0.618252	−0.614218	−0.617806	–
	5000	0.577366	0.577222	0.576750	1.387179	1.379640	1.340894	−0.713591	−0.671955	−0.666950	–
	10000	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 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 of the estimates and the asymptotic covariance matrix 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 for different sample sizes n with $N = 5000$. The line, preceded by ∞, gives 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	
	1000	0.577343	0.577049	0.577455	0.081122	0.076944	0.076627	−0.040666	−0.040421	−0.036237	–
	5000	0.577349	0.577424	0.577241	0.071714	0.071565	0.070109	−0.036610	−0.035116	−0.034965	–
	10000	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
	1000	0.576909	0.577237	0.577743	0.062691	0.062724	0.061552	−0.031817	−0.030782	−0.030871	0.017500
	5000	0.577193	0.577334	0.577497	0.049946	0.051876	0.050206	−0.025787	−0.024146	−0.026080	0.013040
	10000	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
	1000	0.577607	0.576964	0.577018	0.176862	0.177429	0.178333	−0.087992	−0.089192	−0.089015	0.147942
	5000	0.577329	0.577313	0.577328	0.155366	0.157960	0.154613	−0.079387	−0.075988	−0.078577	0.124000
	10000	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
	1000	0.577425	0.577094	0.577126	0.156489	0.156771	0.155526	−0.078799	−0.077816	−0.077693	0.124233
	5000	0.577316	0.577328	0.577332	0.143959	0.145212	0.143865	−0.072627	−0.071307	−0.072569	0.112765
	10000	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	–
	1000	0.577145	0.577250	0.576976	0.268550	0.261884	0.255396	−0.137415	−0.131186	−0.124071	–
	5000	0.577358	0.577305	0.577239	0.279751	0.298353	0.280200	−0.148836	−0.130687	−0.149572	–
	10000	0.577360	0.577378	0.577235	0.295791	0.293497	0.308889	−0.140208	−0.155558	−0.153289	–
	∞	0.57735	0.57735	0.57735	–	–	–	–	–	–	

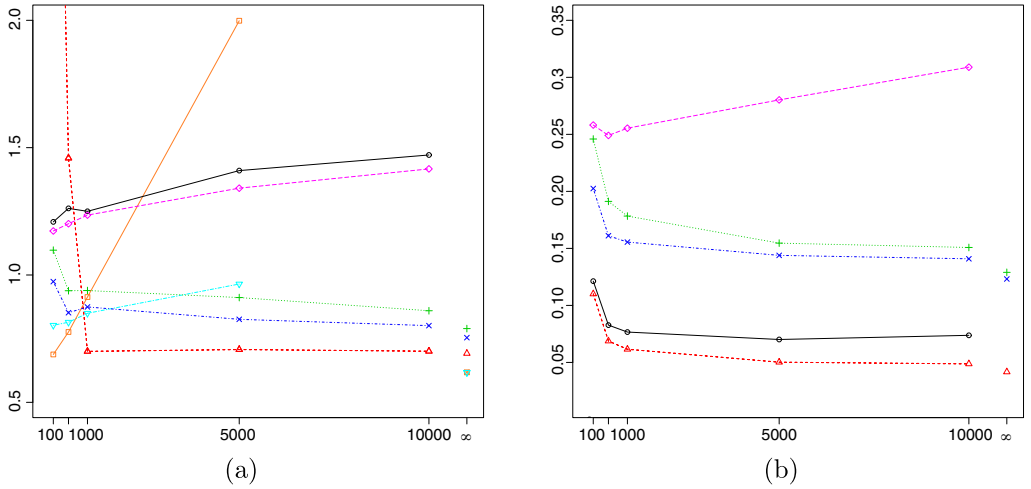

FIGURE 6 n times the variance of α_3 as a function of the sample size n for the simulation model (a) with $X_i \sim U[0,1]$ and (b) with $X_i \sim N(0,1)$ for the LSE (solid, black, ∘), SSE (dashed, red, △), MRCE (dotted, green, +), MRE (dashed-dotted, blue, ×), LS-MRE (long-dashed, pink, ⋄), EDF (two-dashed, light blue, ▽), and PLSE (solid, orange, □). The point at ∞ represents the asymptotic values

Therefore, we do not report results for the sample size $n = 10000$ and simulated only $N = 2500$ data sets for the PLSE with $n = 5000$. Boxplots of $\sum_{j=0}^{3}(\hat{\alpha}_j - \alpha_{0j})^2/3$, shown in Figure 7, illustrate that the PLSE and EFM estimates 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 EDF estimates depend furthermore on smoothing parameters that need to be selected carefully. Figure 6 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 involving 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 an $\sqrt{n}$-consistent estimate, its performance remains inferior to the score procedure in Section 2.2. Nevertheless, it remains an interesting topic to understand the behavior of the LSE in the monotone single-index model.

4 | REAL DATA EXAMPLE

In this section, we apply the estimation techniques to the ozone data (Chambers, Cleveland, Kleiner, & Tukey, 1983). The data set contains observations on the ozone concentration for

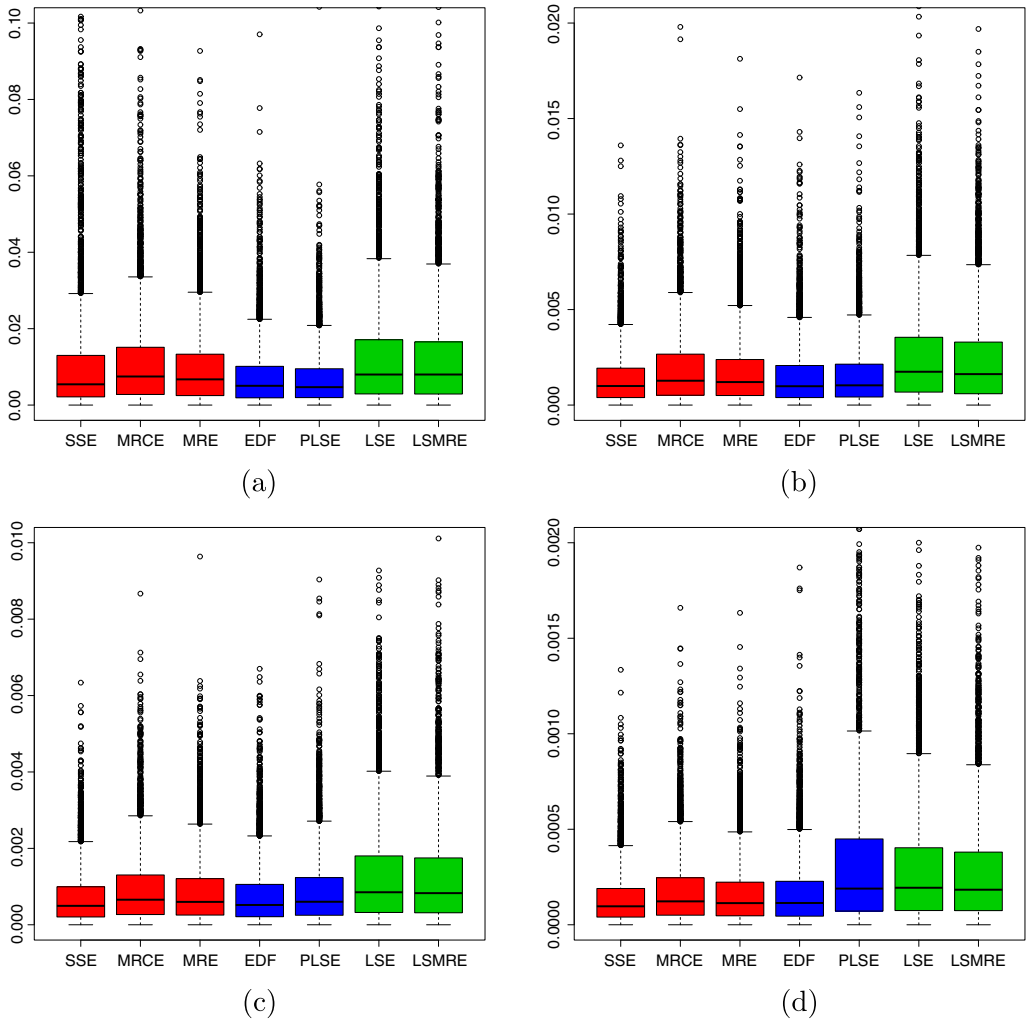

FIGURE 7 Boxplots of $\sum_{j=0}^{3}(\hat{\alpha}_j - \alpha_{0j})^2/3$ for the model with $X_i \sim U[0,1]$ for sample sizes (a) $n = 100$, (b) $n = 500$, (c) $n = 1000$, and (d) $n = 5000$. Red boxes correspond to $\sqrt{n}$-consistent but ineffcient methods (SSE, MRCE, and MRE), blue boxes correspond to $\sqrt{n}$-consistent and efficient methods (EDF and PLSE), and green boxes correspond to methods with an unknown limiting distribution (LSE and LS-MRE)

153 consecutive days between May 1, 1973 and September 30, 1973. We study the relationship between the ozone concentration (Y, ppb) and the meteorological variables, namely, solar radiation (R, Ly), temperature (T, °F), and wind speed (W, mph), in a subset of the data consisting of 111 complete observations.

Yu and Ruppert (2002) compared linear, semiparametric, and fully nonparametric fits for the ozone data and concluded that better results were obtained with a single-index model. The data set became a benchmark in the study of single-index models (Alquier & Biau, 2013; Antoniadis et al., 2004; Karabatsos, 2009; and Wang, 2009, among others), and the results, previously presented in the statistical literature, suggest that a monotone fit for the underlying link function is plausible for the ozone data.

For our data analysis, we have scaled the covariates to have mean 0 and variance 1. Table 5 summarizes the results of the regression parameters for the LSE, SSE, MRCE, MRE, and LS-MRE.

TABLE 5 Ozone data

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

Note. LSE = least squares estimator; SSE = simple score estimator; MRCE = maximum rank correlation estimator; MRE = maximum rank estimator.

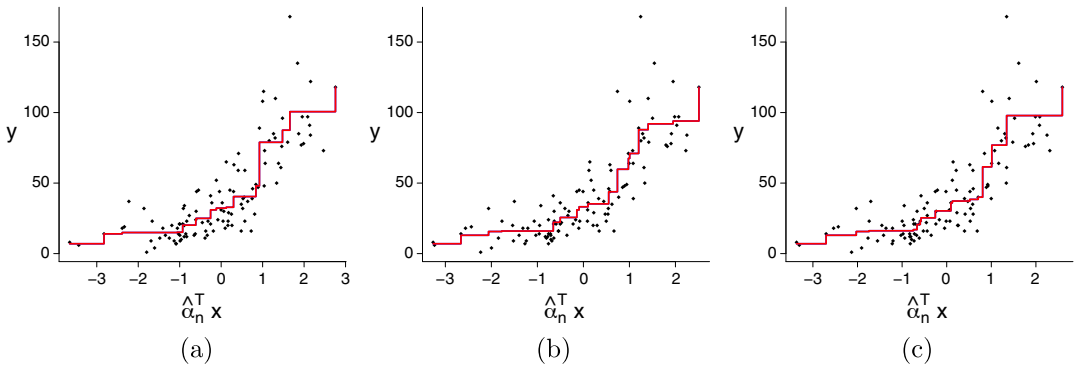

FIGURE 8 Ozone data. Scatterplot $(\hat{\boldsymbol{\alpha}}_n^T \boldsymbol{x}_i, y_i)$ and $\hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}$ (red, step function) for (a) LSE, (b) SSE, and (c) LS-MRE

The estimate $\hat{\psi}_{n\hat{\boldsymbol{\alpha}}_n}$ of ψ_0, together with a scatterplot of $(\hat{\boldsymbol{\alpha}}_n^T \boldsymbol{x}_i, y_i)$, is given in Figure 8 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.

ACKNOWLEDGEMENTS

The research of the second author was supported by the Research Foundation Flanders (FWO) under Grant 11W7315N. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged. For the simulations, we used the infrastructure of the VSC—Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government—Department EWI.

ORCID

Piet Groeneboom http://orcid.org/0000-0001-8027-8114
Kim Hendrickx http://orcid.org/0000-0003-4005-2676

REFERENCES

Alquier, P., & Biau, G. (2013). Sparse single-index model. *Journal of Machine Learning Research*, *14*, 243–280.

Antoniadis, A., Grégoire, G., & McKeague, I. W. (2004). Bayesian estimation in single-index models. *Statistica Sinica*, 1147–1164.

Aragón, J., & Quiroz, A. J. (1995). Rank regression for current status data. *Statististics & Probability Letters*, *24*(3), 251–256.

Balabdaoui, F., Durot, C., & Jankowski, H. (2016). Least squares estimation in the monotone single index model. arXiv preprint arXiv:1610.06026.

Balabdaoui, F., Groeneboom, P., & Hendrickx, K. (2017). Score estimation in the monotone single index model. arXiv Preprint arXiv:1712.05593.

Barlow, R., Bartholomew, D., Bremner, J., & Brunk, H. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression*, Wiley Series in Probability and Mathematical Statistics. London–New York–Sydney: John Wiley & Sons.

Cavanagh, C., & Sherman, R. P. (1998). Rank estimators for monotonic index models. *Journal of Econometrics*, *84*(2), 351–381.

Chambers, J., Cleveland, W., Kleiner, B., & Tukey, P. (1983). *Graphical methods for data analysis*. Boston: Duxbury Press.

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.

Cui, X., Härdle, W. K., & Zhu, L. (2011). The EFM approach for single-index models. *The Annals of Statistics*, *39*(3), 1658–1688.

Delecroix, M., Härdle, W., & Hristache, M. (2003). Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, *86*(2), 213–226.

Groeneboom, P., & Hendrickx, K. (2017). Current status linear regression. *The Annals of Statistics*. Retrieved from https://arxiv.org/abs/1601.00202

Groeneboom, P., & Jongbloed, G. (2014). *Nonparametric estimation under shape constraints*. Cambridge: Cambridge University Press.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, *35*(2–3), 303–316.

Härdle, W., Hall, P., & Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, *21*(1), 157–178.

Hooke, R., & Jeeves, T. A. (1961). "Direct search" solution of numerical and statistical problems. *Journal of the ACM*, *8*(2), 212–229.

Hristache, M., Juditsky, A., & Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *The Annals of Statistics*, 595–623.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*, *24*(2), 540–568.

Huang, J., & Wellner, J. (1993). Regression models with interval censoring. *Proceedings of the Kolmogorov Seminar, Euler Mathematics Institute*, St. Petersburg, Russia.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, *58*(1–2), 71–120.

Karabatsos, G. (2009). Modeling heteroscedasticity in the single-index model with the Dirichlet process. *Advances and Applications in Statistical Sciences*, *1*(1), 83–104.

Klein, R. W., & Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, *61*(2), 387–421.

Kuchibhotla, A. K., & Patra, R. K. (2017). Efficient estimation in convex single index models. Retrieved from https://arxiv.org/abs/1708.00145

Li, G., & Zhang, C.-H. (1998). Linear regression with interval censored data. *The Annals of Statistics*, *26*(4), 1306–1327.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, *3*(3), 205–228.

Murphy, S. A., van der Vaart, A. W., & Wellner, J. A. (1999). Current status regression. *Mathematical Methods of Statistics*, *8*(3), 407–425.

Newey, W. K., & Stoker, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica: Journal of the Econometric Society*, 1199–1223.

Robertson, T., Wright, F., & Dykstra, R. (1988). *Order restricted statistical inference*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Chichester: John Wiley & Sons.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica*, *61*(1), 123–137.

Torczon, V. (1997). On the convergence of pattern search algorithms. *SIAM Journal on Optimization*, *7*(1), 1–25.

Wang, H.-B. (2009). Bayesian estimation and variable selection for single index models. *Computational Statistics & Data Analysis*, *53*(7), 2617–2627.

Xia, Y., & Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, *97*(5), 1162–1184.

Yu, Y., & Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, *97*(460), 1042–1054.

How to cite this article: Groeneboom P, Hendrickx K. Estimation in monotone single-index models. *Statistica Neerlandica*. 2019;73:78–99. https://doi.org/10.1111/stan.12138

APPENDIX

A.1 | Equivalence MRCE and MRE for the binary choice model and the current status regression model

In this section, we show that the maximizers of H_n and R_n coincide for the current status linear regression model. A similar derivation holds for the binary choice model.

Consider the MRE for the current status model maximizing

$$R_n(\boldsymbol{\alpha}) = \sum_{i \neq j} \Delta_i \left\{ T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i \beta > T_j - \boldsymbol{\alpha}^T \boldsymbol{X}_j \right\}.$$

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

$$\begin{aligned} H_n(\boldsymbol{\alpha}) &= \sum_{i \neq j} \{\Delta_i > \Delta_j\} \left\{ T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i \beta > T_j - \boldsymbol{\alpha}^T \boldsymbol{X}_j \right\} \\ &= \sum_{i \neq j} \Delta_i (1 - \Delta_j) \left\{ T_i - \boldsymbol{\alpha}^T \boldsymbol{X}_i \beta > T_j - \boldsymbol{\alpha}^T \boldsymbol{X}_j \right\}. \end{aligned}$$

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

$$\begin{aligned} H_n(\boldsymbol{\alpha}) &= \sum_{j<i} \Delta_i (1 - \Delta_j) = \sum_{j<i} \Delta_i - \sum_{j<i} \Delta_i \Delta_j = \sum_{j<i} \Delta_i - \frac{1}{2} \sum_{j \neq i} \Delta_i \Delta_j \\ &= R_n(\boldsymbol{\alpha}) - \frac{1}{2} \sum_{j \neq i} \Delta_i \Delta_j. \end{aligned}$$

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 $R_n(\boldsymbol{\alpha})$ and $H_n(\boldsymbol{\alpha})$ coincide and both estimators are equivalent.

A.2 | Asymptotic distribution of the MRCE and the MRE

In this section, we show that the asymptotic normal distribution for the MRCE is given by

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

where $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ are defined in Table 1 and where $\boldsymbol{V}^-$ denotes the Moore–Penrose inverse of matrix $\boldsymbol{V}$. 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 is derived in Sherman (1993) and Cavanagh and Sherman (1998), where the authors restrict the parameter space to a compact subset of $\{\boldsymbol{\alpha} \in \mathbb{R}^d : \alpha_d = 1\}$. Each $\boldsymbol{\alpha}$ is represented as $(\theta, 1)$, 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 Balabdaoui et al. (2017), it follows, by similar arguments as in Sherman, that for the MRCE, we have

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

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

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

converges in distribution to a normal $N_{d-1}(\mathbf{0}, \boldsymbol{W})$ random vector. Here, ∇_1 represents the first partial derivative operator with respect to $\boldsymbol{\beta}$. Let $\boldsymbol{A}$ denote $\mathbb{E}(\nabla_2 \tau(\boldsymbol{X}, Y), \boldsymbol{\beta})$. If $\boldsymbol{A}$ is negative definite, then it follows, by theorem 2 in Sherman, that

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

Using an application of the delta method, we conclude that

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

where $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ are given in Table 1 and where the last equality follows analogously to the proof of asymptotic normality given in Balabdaoui et al. (2017).