DOCTORAATSPROEFSCHRIFT

2010 | Faculteit Wetenschappen

The Koziol-Green and Generalized Koziol-Green Model with Covariates under Dependent Censoring

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in Wetenschappen, Wiskunde, te verdedigen door:

Auguste GADDAH

Promotor: prof. dr. Roel Braekers

universiteit hasselt

www.uhasselt.be

Universiteit Hasselt | Campus Diepenbeek
Agoralaan | Gebouw D | BE-3590 Diepenbeek | België
Tel.: +32(0)11 26 81 11

DOCTORAATSPROEFSCHRIFT

2010 | Faculteit Wetenschappen

The Koziol-Green and Generalized Koziol-Green Model with Covariates under Dependent Censoring

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in Wetenschappen, Wiskunde, te verdedigen door:

Auguste GADDAH

Promotor: prof. dr. Roel Braekers

D/2010/2451/28

universiteit hasselt

Voor mijn moeder

en

ter nagedachtenis van mijn vader

Acknowledgements

Behind every achievement, lies at least one ”spring board”. This thesis is not an exception and I am wracked up in huge debts. First and foremost, I do acknowledge my promotor Prof. dr. Roel Braekers for meticulously supervising my PhD work from its inception to completion. I heartedly do appreciate his patience, guidance, encouragement and time spent on me. Also, I would like to express my profound gratitude and thankfulness to the jury members: Prof. dr. Gerda Claeskens, Katholieke Universiteit Leuven; Prof. dr. Ingrid van Keilegom, Université Catholique de Louvain; Prof. dr. Jacobo de Uña-Álvarez, Universidade de Vigo; Prof. dr. Nöel Veraverbeke, Universiteit Hasselt and Prof. dr. Paul Janssen, Universiteit Hasselt. Their constructive comments and suggestions have tremendously improved both the content and style of the thesis.

I am grateful to Tetyana Kadankova, Tom Jacobs, Yves Grouwels and Tim Willaert. They made my working environment a very conducive abode. It has been a pleasure to share the office with you. The lunch hours I shared with Saskia Litière, Jose Cortinas and Ariel Alonso are not forgotten. Their cheerful ”words” not only served as a pastime from the morning’s mathematics, but also a rejuvenation for the afternoon’s mathematics. In addition, I do express my thankfulness to Carolyn Collinet, Danny Dendas, Ruud Van Thienen, Davy Nijs, Manu Houberigts, Alexander Jageneau and the rest of Wado Ruy Karateclub members. The sessions we had were very effective and immeasurably contributed to my research aptitude.

Finally, I extend my gratitude to the whole I-BioStat staff for their distinguishable support over these years. To my family, friends and all those who contributed in one way or another, I say THANK YOU. With candid gratefulness, I simply say ”Eh-Sheh-Wuh” to Adetayo Kasim and his wife Ebunola Kasim.

Auguste GADDAH
Diepenbeek, September 2010

Contents

1

Introduction

The study of non-negative response variables is crucial and takes several forms in a wide variety of areas of modern scientific investigations. One of these is lifetime or survival time studies, where the response variable is expressed as the time until certain event of interest (time-to-event endpoint). In engineering for example, researchers are often interested in studying the time until the break down of a machine component. Another example is in the social sciences, where interest lies in the duration of strikes, duration of unemployment or the duration of marriages in societies. In medical settings, survival times emerge from investigations that focus on the time until recurrence of cancer tumors, the time to recovery after a surgical operation or the life span of some biological units, among others. Nonetheless, there are also cases in survival time studies where the

term ”time” may not represent the literal time. For instance, in quality control or reliability in manufacturing, this could be the amount of force needed to render a part unusable. While in economics, it could also be the amount payed by an insurance company in case of damage.

In various fields of survival time studies, researchers are often confronted with the distinguishable and unifying phenomenon of censoring. This surfaces as a consequence of the fact that, for some study units, the exact survival time is known, whereas for others only a partial information is available. Censoring in general occurs for various reasons. Depending on the underlying reason for censoring, we can broadly distinguish between three types of censoring, namely *Type I*, *Type II* and *Random* censoring schemes. Type I censoring occurs when the censoring time is fixed *a priori*. While in type II censoring, the censoring time is determined by a fixed number of exact survival times to be observed. In both these types of censoring however, the censoring mechanism is controlled by the investigator. In a laboratory experiment for example, a researcher who wants to investigate the lifespan of a number of fluorescent tubes may put them on a test in order to record their times to failure. Some tubes may take a long time to burn out and it may not be feasible for the experimenter to wait that long. Therefore, he/she may decide to end the experiment at a prescribed time (i.e. fixed censoring time). In such situation, the exact lifetime of some tubes may not be observed and this leads to Type I censoring. On the other hand, the investigator may not have a prior knowledge of the appropriate fixed censoring time and may chose to wait until a prespecified proportion of the tubes burns out. The exact lifetime of some tubes may not be observed in this second scenario as well, in which case we have type II censoring. Obviously, the censoring mechanisms in these scenarios are under the control of the investigator.

Random censoring on the other hand, is beyond the control of the investigator. It occurs when the response random variable of each study unit is associated with a potential censoring random variable. Thus, in a study where the lifetime of primary interest is the time until death from a heart disease, it is possible that some study units would die from other diseases and their exact lifetime cannot be observed. Also, patients with inoperable cancer are often taken off study when their tumor grows in size by a certain amount or when new lesions are detected and as such their exact lifetime cannot be

observed. Clearly, the censoring variable (i.e. death from other diseases or taken off study) is random and cannot be controlled by the researcher.

Between the aforementioned censoring schemes, random censoring is the most predominant and can further be discerned as three types. The first is *random left censoring* and occurs when the available partial information is an upper bound on response of interest. To illustrate this, consider the African children example of Miller (1981) where interest is on knowing the age at which certain group of children learn to perform a certain task. At the beginning of the research, some children already knew how to perform the task. In such cases, the only available information is that those children can perform the task at a younger age. Thus, the age at which those children knew the task is left censored at their respective current age.

The second is *random right censoring*, where the available partial information is a lower bound on the response. In the heart disease example, if a patient dies without a heart disease then the only information on his time to heart disease is that this time is greater than the observed death time and as such is right censored. Both these types of random censoring can be considered as special cases of the third type called *interval censoring*. In this latter type of censoring, the available partial information is that the response time of interest falls within a certain interval. This is the case for example in a HIV-AIDS study, where the study subjects are examined yearly for HIV infection. Therefore, if a subject is not infected at year 3 but found to be infected at year 4, then the only information on the time of infection is that it is less than 4 years but greater than 3 years. As a result, the infection time of that subject is interval censored between year 3 and 4.

Although censoring is regarded as nuisance, it is often an integral component of most survival studies. As a result, the statistical analysis of such data sets requires the use of special techniques. Another characterizing feature of survival studies is the availability of some additionally measured variables (covariates). These covariates in most cases are not of primary interest to the researcher, but have the potential to influence the distribution of the time until the event of interest. In other words, the distribution of the lifetime variables varies with different covariate values. As an example, imagine a study that is supposed to provide insight into the distribution of the length of stay of patients in hospital admission. Then it is apparent that the distribution of interest may be influenced

by the age of the patient and/or the severity of his/her medical condition.

In this thesis, we provide some new techniques that are associated with the statistical analysis of censored survival data. Primarily, we focus on generalizations of the random right censorship Koziol-green model in the absence and presence of covariates. For easy exposition, we first give a more rigorous introduction of the setting without covariates in Section 1.1. In that same section, we give a brief review of some basic and existing statistical techniques that are usually employed in this respect. Afterwards, we introduce our new extension of the Koziol-Green model in Subsection 1.1.1 for the case without covariates. In Section 1.2, we vividly describe the setting whereby some covariates which are thought to contain some information about the lifetime of interest are collected together with the censored responses. There, we introduce the conditional Koziol-Green model and its new generalization thereof. Before proceeding, it is important to note that the representation of censored data in Section 1.2 is based on fixed design (covariate) points. Nevertheless, the associated methodologies can be applied to the random design settings, with some modifications.

1.1 Random right censored survival time

Suppose $Y_1, Y_2, Y_3, ..., Y_n$ is a sample of n independent identically distributed non-negative response variables with a continuous distribution function $F(t) = P(Y_1 \leq t)$. Frequently, these responses are subject to random right censoring. That is, for every $Y_i \sim Y$ $(i = 1, 2, 3, ..., n)$, there exist a potential non-negative random variable $C_i \sim C$, called censoring variable with distribution function $G(t) = P(C_1 \leq t)$ such that we can only observe $Z_i = \min(Y_i, C_i)$ and $\delta_i = \mathbb{1}\{Y_i \leq C_i\}$ where the couples (Z_i, δ_i) are independent copies of (Z, δ). Let us denote the distribution of the the Z-sample by $H(t) = P(Z \leq t)$ and assume that Y and C are independent. Then we can write

$$1 - H(t) \quad = \quad (1 - F(t))(1 - G(t)).$$

Under this assumption, the well known Kaplan and Meier (1958) product limit estimator for the distribution function F serves as the inferential bases for the lifetime of interest

and is given by

$$F_n^{KM}(t) \quad = \quad 1-\left\{\prod_{i:Z_i\leq t}\left(\frac{n-R_i}{n-R_i+1}\right)^{\delta_i}\right\},$$

where R_i is the rank of the ith observation in the Z-sample. This estimator is a step function which jumps only at the uncensored observations. In the absence of censoring, it is easy to see that this estimator reduces to the empirical distribution function

$$F_n(t) \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{Y_i\leq t\}.$$

In some settings however, the censoring variable is informative to the lifetime variable through its distribution function. In such case, Koziol and Green (1976) considered a sub-model under the assumption that the survival time variable Y and the censoring time variable C are independent. In the Koziol-Green sub-model, the authors assumed that the distribution function of the censoring variable is a power of the distribution of the lifetime variable. Mathematically, this is expressed as

$$1-G(t) \quad = \quad (1-F(t))^{\beta} \quad , \quad \forall t\geq 0 \tag{1.1}$$

for some $\beta>0$. This Koziol and Green (1976) sub-model is equivalent to the extra assumption that the observable time Z and censoring indicator δ are mutually independent (Sethuraman (1965), Kochar and Proschan (1991)). This means that, the instantaneous event rate is proportional to the instantaneous rate of censoring. This extra assumption, introduced by Koziol and Green (1976) not only allowed the censored observations to contribute to the estimation of survival distribution function of interest, but also marked the era of informative censoring within the domain of lifetime analysis whereby the censoring distribution is allowed to depend on unknown parameters of the lifetime distribution. In light of this, the Koziol and Green (1976) model received considerable attention in the statistical literature. For example, Abdushukurov (1987) and Cheng and Lin (1987) independently derived the model

$$F^{ACL}(t) \quad = \quad 1-(1-H(t))^{\gamma}$$

for the survival time distribution function under the Koziol and Green (1976) characterization, where $\gamma=\frac{1}{1+\beta}=P(\delta=1)$ is the expected proportion of uncensored observations

and $H(t) = P(Z \leq t)$ is as defined earlier. Upon replacing γ and $H(\cdot)$ by appropriate estimators, these authors obtained a non-parametric maximum likelihood estimator for the survival time distribution function under the Koziol-Green model and later studied its large sample properties. The authors also showed that the non-parametric maximum likelihood estimator under the Koziol and Green (1976) model is asymptotically more efficient than the corresponding product limit estimator of Kaplan and Meier (1958). The estimator studied by Abdushukurov (1987), Cheng and Lin (1987) is of the form

$$F_n^{ACL}(t) \quad = \quad 1-(1-H_n(t))^{\gamma_n} \tag{1.2}$$

where γ_n is the proportion of uncensored observations and $H_n(t)$ is the empirical distribution of the observed time, which are respectively given by

$$\begin{aligned} \gamma_n &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\delta_i = 1\}, \\ H_n(t) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Z_i \leq t\}. \end{aligned} \tag{1.3}$$

For a detailed review of the statistical literature on Koziol and Green (1976) model, we refer to Csörgó (1988). He also developed a test to check the validity of the Koziol-Green model and showed that there are many data sets for which (1.1) is not satisfied. Equivalently, there are many practical settings for which the observable time Z and censoring indicator δ are not independent. For those cases, it may be questionable to use the estimator (1.2) as the inferential basis for the lifetime. The first contribution of this thesis is to derive and study an extended estimator which has the ability to overcome the limitations of (1.2). We introduce such estimator in Section 1.1.1 and later proceed with a detailed study of its properties and applicability in Chapters 2 and 3.

1.1.1 The extended Koziol-Green model under dependent censoring

As mentioned above, the applicability of the estimator (1.2) proposed independently by Abdushukurov (1987) and Cheng and Lin (1987) under the classical Koziol-Green model (1.1) could be limited in practice, due to the independence assumption that it imposes on the observable time Z and the censoring indicator δ (Csörgó (1988)). Here,

we ameliorate this limitation and assume instead, the existence of some known copula function $\mathscr{C}$ that describe the joint distribution of Z and δ. Mathematically, we express this as

$$H^u(t) \quad = \quad P(\delta = 1, Z \leq t) = \mathscr{C}(\gamma, H(t)), \tag{1.4}$$

where γ and $H(\cdot)$ are as previously defined. In the parlance of Klement et al. (2007), $\mathscr{C}(\gamma, H(t))$ is known as the vertical γ-section of the copula $\mathscr{C}$, for a fixed $\gamma \in (0,1)$. There, the set of all copulas with the same vertical γ-section were studied. The authors also found copulas that bound both below and above the set. Because the censoring indicator δ is a discrete variable, we also know from Sklar's theorem (see Nelsen (2006)) that $\mathscr{C}$ may not be unique. As a result, it is important to be conscious and not directly interpret (1.4) as a dependence model, but rather as a device to help relax the Koziol and Green (1976) characterization (1.1). Furthermore from Genest and Nešlehová (2007), it is also clear that the copula function $\mathscr{C}$ alone is not sufficient to describe the association structure between Z and δ. The marginal distributions are also needed for this. However, it turns out that the non-uniqueness of the copula function $\mathscr{C}$ does not have any significant practical consequence on the estimator for the distribution function F of the survival time. In the sequel, we will see that this estimator does not change when a copula function with the same vertical γ-section is chosen. In passing, note that $\gamma = 0$ corresponds to the situation with only censored observations, in which case it is not feasible to make inference about the survival time. When $\gamma = 1$, then we have fully observed lifetimes and it is not necessary to account for censoring in order to make inference about the lifetime distribution. Thus, it is reasonable to assume that $\gamma \in (0,1)$.

On the other hand, it is imperative to make a non-verifiable assumption about the relationship between survival time Y and the censoring time C in order to proceed and derive an estimator for the marginal distribution function of the survival time (Tsiatis (1975)). It is common in time to event analysis to assume independence between these random variables. In some situations however, this assumption may be doubtful and unrealistic. For example, in a cancer study where the event of interest is the recurrence of a cancer tumor and the censoring event is death, or in industrial testing, it may occur that a piece of equipment is taken away (i.e. censored) because it shows signs of future failure. Adopting the strategy of some previous authors (e.g. Zheng and Klein (1995),

Rivest and Wells (2001), Braekers and Veraverbeke (2005)), we solve this problem by using a copula model to describe the possible dependence structure of Y and C. In order to obtain tractable results, we only concentrate on the class of Archimedean copulas to model the joint distribution of Y and C. That is,

$$S(t_1,t_2) \quad = \quad P(Y>t_1, C>t_2) = \varphi^{[-1]}\left(\varphi(\bar{F}(t_1)) + \varphi\left(\bar{G}(t_2)\right)\right) \tag{1.5}$$

where $\bar{F}(t) = 1-F(t)$ and $\bar{G}(t) = 1-G(t)$ are survival distribution functions of Y and C respectively, $\varphi : [0,1] \to [0,\infty]$ is a known continuous, convex and strictly decreasing function with $\varphi(1) = 0$. We denote by $\varphi^{[-1]}$ the pseudo-inverse of φ which is defined as in Nelsen (2006),

$$\varphi^{[-1]} \quad = \quad \begin{cases} \varphi^{-1}(s) & , \quad 0 \leq s \leq \varphi(0) \\ 0 & , \quad \varphi(0) \leq s \leq \infty \end{cases}.$$

Using relation (1.5), we now derive an estimator for the distribution function F under model (1.4). To do so, we work in parallel with Tsiatis (1975) and obtain from (1.5) that

$$\frac{dH^u(t)}{dt} \quad = \quad -\frac{\partial}{\partial t_1} S(t_1,t_2)\Big|_{t=t_1=t_2} = \frac{\varphi'(\bar{F}(t))}{\varphi'(S(t,t))}\frac{dF(t)}{dt} = \frac{\varphi'(\bar{F}(t))}{\varphi'(H(t))}\frac{dF(t)}{dt},$$

with $\varphi'(u) = \frac{d}{du}\varphi(u)$ and $S(t,t) = \varphi^{-1}\left(\varphi(\bar{F}(t)) + \varphi(\bar{G}(t))\right) = 1 - H(t) = \bar{H}(t)$.

Reorganizing this equation, gives

$$\varphi'(\bar{F}(t))\frac{dF(t)}{dt} \quad = \quad \varphi'(\bar{H}(t))\frac{dH^u(t)}{dt}.$$

By integrating on both sides and with $\varphi(\bar{F}(0)) = \varphi(1) = 0$, we obtain that

$$\bar{F}(t) \quad = \quad \varphi^{-1}\left(-\int_0^t \varphi'(\bar{H}(s))dH^u(s)\right). \tag{1.6}$$

From the informative censoring structure described in the extended Koziol-Green model and given by (1.4), we find that

$$dH^u(s) \quad = \quad \mathscr{C}_{01}(\gamma, H(s))\, dH(s)$$

where $\mathscr{C}_{01}(u,v) = \frac{\partial}{\partial v}\mathscr{C}(u,v)$ is the first partial derivative of the general copula function $\mathscr{C}(u,v)$ with respect to the second coordinate. Introducing this later relation into (1.6) in conjunction with a variable transformation, we obtain the model

$$\bar{F}(t) \quad = \quad \varphi^{-1}\left(-\int_0^{H(t)} \varphi'(1-w)\mathscr{C}_{01}(\gamma, w)\, dw\right),$$

in which we allow for dependent censoring as well as non-independency of Z and δ. In the above model, we note that the copula function $\mathscr{C}$ only enters through its derivative $\mathscr{C}_{01}$ which is a conditional probability (see for example Nelsen (2006, page 41)). When the survival time is independent of the censoring time, then this model connects to the ideas of some previous authors through the relation

$$P(\delta = 1|Z = t) \quad = \quad \mathscr{C}_{01}(\gamma, H(t))$$

of the conditional probability of an uncensored observation, given the observed lifetime. For instance, in the semiparametric random censorship model of Dikta (1998), the author assumed a parametric model for the above conditional probability. While Cao et al. (2005), used a non-parametric kernel smoother for this same conditional probability.

To find the estimator

$$\bar{F}_n(t) \quad = \quad \varphi^{-1}\left(-\int_0^{H_n(t)} \varphi'(1-w)\mathscr{C}_{01}(\gamma_n, w)\, dw\right) \tag{1.7}$$

for the survival distribution $\bar{F}(t)$ in the extended Koziol-Green model, we replaced γ and $H(t)$ by their empirical counterparts which are as defined in (1.3).

If we take $\mathscr{C}$ such that $H^u(t) = \gamma H(t)$, we easily see that this estimator simplifies to

$$\bar{F}_n(t) \quad = \quad \varphi^{-1}(\gamma_n \varphi(\bar{H}_n(t)))\,, \tag{1.8}$$

which is the unconditional version of the conditional estimator of Braekers and Veraverbeke (2008). Moreover, if we also assume that the censoring time and the survival time are independent, then (1.8) reduces to the estimator of Abdushukurov (1987) and Cheng and Lin (1987) as displayed in (1.2). As a result, we obviously see that the estimator (1.7) is more general and includes (1.8) and (1.2) as special cases (se also Table 1.1).

In Chapter 2, we pursue further the estimator (1.7) and obtain some desirable theoretical results. In addition, we present a goodness-of-fit test to determine the validity of this estimator in practical applications in Chapter 3. These results can also be found in Gaddah and Braekers (2010a,b).

1.2 Random right censored survival time in fixed design

In the previous section, we meticulously described censored survival data in the absence of covariate information and later introduced the extended Koziol-Green model which can be used for that setting. The purpose of the current section is to also give a detailed description of the setting with covariates. Let $Y_1, Y_2, Y_3, ..., Y_n$ denote independent responses observed at fixed design points $0 \leq x_1 \leq x_2 \leq x_3 \leq ... \leq x_n = 1$. Similar to the previous setting (i.e. scenario without covariates), it often occurs that these responses are subject to random right censoring. In other words, there exist at each design point x_i, a non-negative random variable C_i such that the observable variables are $Z_i = \min(Y_i, C_i)$ and $\delta_i = \mathbb{1}\{Y_i \leq C_i\}$. Also, let $F_{x_i}(t) = P(Y_i \leq t), G_{x_i}(t) = P(C_i \leq t)$ and $H_{x_i}(t) = P(Z_i \leq t)$ denote the distribution functions of Y_i, C_i and Z_i respectively. At a fixed design point $x \in [0,1]$, we further write F_x, G_x and H_x for the distribution functions of the response variable Y_x at x, the censoring variable C_x at x and the observable time $Z_x = \min(Y_x, C_x)$ at x. It is important to note that we write Y_i, C_i, Z_i, δ_i instead of $Y_{x_i}, C_{x_i}, Z_{x_i}, \delta_{x_i}$ for the design points x_i. If we assume that Y_i and C_i are conditionally independent, given the covariate x_i, then we can write

$$1 - H_{x_i}(t) \quad = \quad (1 - F_{x_i}(t))(1 - G_{x_i}(t)). \tag{1.9}$$

In analogy with the classical Koziol-Green model (1.1), we can also assume that

$$1 - G_x(t) \quad = \quad (1 - F_x(t))^{\beta_x}, \qquad x \in [0,1] \tag{1.10}$$

where $\beta_x > 0$ is allowed to depend only on x. Using assumption (1.9) and model (1.10), subsequently leads to the estimator

$$F_{xh}^{VC}(t) \quad = \quad 1 - (1 - H_{xh}(t))^{\gamma_{xh}} \tag{1.11}$$

in the conditional Koziol-Green model (1.10), where

$$H_{xh}(t) = \sum_{i=1}^{n} w_{n_i}(x, h_n) \mathbb{1}\{Z_i \leq t\} \qquad \text{and} \qquad \gamma_{xh} = \sum_{i=1}^{n} w_{n_i}(x, h_n) \mathbb{1}\{\delta_i = 1\} \tag{1.12}$$

are the Stone (1977) type estimators for $H_x(t)$ and $\gamma_x = \frac{1}{1+\beta_x} = P(\delta_x = 1)$ respectively, with $w_{n_i}(x, h_n)$ being the Gasser-Müller type weight functions based on the kernel K and

defined by

$$\begin{aligned} w_{n_i}(x,h_n) &= \frac{1}{c_n(x,h_n)} \int_{x_{i-1}}^{x_i} \frac{1}{h_n} K\left(\frac{x-z}{h_n}\right) dz, \quad i = 1,2,...,n \\ c_n(x,h_n) &= \int_0^{x_n} \frac{1}{h_n} K\left(\frac{x-z}{h_n}\right) dz \end{aligned}$$

depending on a positive bandwidth sequence $\{h_n\}$, which tends to zero as $n \to +\infty$. The estimator (1.11) was introduced and studied by Veraverbeke and Cadarso-Suárez (2000). It is the conditional version of the one proposed by Abdushukurov (1987) and Cheng and Lin (1987) under the classical Koziol-Green model.

Recently, Braekers and Veraverbeke (2008) further studied the estimator (1.11) under the conditional Koziol-Green model (1.10). More specifically, Braekers and Veraverbeke (2008) extended (1.11) to accommodate possible dependence between the survival time variable Y_x at x and the censoring variable C_x at x. Also, the authors proved the consistency and asymptotic normality of their estimator. In this thesis, we obtain further insight into this recent extension of Braekers and Veraverbeke (2008) under the conditional Koziol-Green model. In particular, we complement their work with a weak convergence result. As an application of the weak convergence, we show the asymptotic efficiency of the conditional estimator of Braekers and Veraverbeke (2008) over the copula graphics estimator of Braekers and Veraverbeke (2005). In addition, we develop a confidence band and illustrate its use on simulated as well as real data set. For convenience, we reintroduce this estimator below and defer the associated new results to Chapter 4.

1.2.1 The conditional Koziol-Green model under dependent censoring

In what preceded, we introduced the general setting of censored data in the presence of covariate information. In analogy with the setting without covariates, we need to make a non-verifiable assumption about the underlying dependence structure that describes the relationship between the lifetime variable and the censoring variable in order to uniquely estimate the marginal distribution function of the lifetime variable. Conditional on the covariate, it is common to assume independence between these variables. However, in

some situations this assumption may not be feasible. For example, in medicine, we are often interested in the time until dying from a certain disease. This time may be related to the time until dying from another disease. Therefore, we again need a dependence model for the association between the time until event of interest and the time until censoring (i.e. death from other diseases). In line with (1.5), we assume that at a fixed design point $x \in [0,1]$, the join survival distribution of the lifetime Y_x and censoring time C_x satisfies

$$S_x(t_1,t_2) \quad = \quad \varphi_x^{[-1]}\left(\varphi_x(\bar{F}_x(t_1)) + \varphi_x\left(\bar{G}_x(t_2)\right)\right) \tag{1.13}$$

where for each $x \in [0,1]$, $\bar{F}_x = 1 - F_x$ and $\bar{G}_x = 1 - G_x$ are the conditional survival distribution functions of the lifetime Y_x and censoring time C_x respectively, $\varphi_x : [0,1] \to [0,+\infty]$ is a known continuous, convex strictly decreasing function with $\varphi_x(1) = 0$. $\varphi_x^{[-1]}$ is the pseudo inverse of φ_x and given by

$$\varphi_x^{[-1]}(s) \quad = \quad \begin{cases} \varphi_x^{-1}(s) & , \quad 0 \le s \le \varphi_x(0) \\ 0 & , \quad \varphi_x(0) \le s \le +\infty \end{cases}$$

Similar to the derivation of the extended estimator (1.7) introduced earlier, it follows from relation (1.13) that

$$\varphi_x'(\bar{F}_x(t))\frac{\partial}{\partial t}F_x(t) \quad = \quad \varphi_x'(\bar{H}_x(t))\frac{\partial}{\partial t}H_x^u(t) \tag{1.14}$$

with $H_x^u(t) = P(Z_x \le t, \delta_x = 1), \varphi_x'(u) = \frac{\partial}{\partial u}\varphi_x(u)$ and

$$\bar{H}_x(t) \quad = \quad 1 - H_x(t) = S_x(t,t) = \varphi_x^{[-1]}\left(\varphi_x(\bar{F}_x(t)) + \varphi_x\left(\bar{G}_x(t)\right)\right).$$

By the extra assumption (1.10), which is equivalent to the conditional independence of Z_x and δ_x, it follows that

$$H_x^u(t) \quad = \quad P(Z_x \le t)P(\delta_x = 1) = H_x(t)\gamma_x.$$

Plugging the preceding display into equation (1.14) and integrating on both sides yields

$$\varphi_x(\bar{F}_x(t)) \quad = \quad \gamma_x\varphi_x(\bar{H}_x(t)) \tag{1.15}$$

On recalling that $\bar{H}_x(t) = \varphi_x^{[-1]}\left(\varphi_x(\bar{F}_x(t)) + \varphi_x\left(\bar{G}_x(t)\right)\right)$ for all $t \ge 0$, the preceding display can be expressed as

$$\bar{G}_x(t) \quad = \quad \varphi_x^{[-1]}\left(\beta_x\varphi_x(\bar{F}_x(t))\right) \quad , \quad \beta_x = \frac{1-\gamma_x}{\gamma_x} \tag{1.16}$$

which is equivalent to relation (1.10) under dependent censoring.

Relations (1.15) and (1.16) are given in Braekers and Veraverbeke (2008). To find an estimator for the conditional distribution function of the lifetime, these authors rewrote relation (1.15) as

$$F_x(t) \quad = \quad 1-\varphi_x^{[-1]}(\gamma_x\varphi_x(\bar{H}_x(t)))$$

and replaced the different unknown quantities γ_x and $H_x(t)$ by the corresponding estimators γ_{xh} and $H_{xh}(t)$ to obtain

$$F_{xh}^{BV}(t) \quad = \quad 1-\varphi_x^{[-1]}(\gamma_{xh}\varphi_x(\bar{H}_{xh}(t))) \tag{1.17}$$

where γ_{xh} and $H_{xh}(t)$ are the Stone (1977) type non-parametric estimators as given in (1.12). In Gaddah and Braekers (2009), we further studied the conditional Koziol-Green model under dependent censoring and proved the weak convergence of the process associated with the estimator (1.17) to a zero mean Gaussian process. In Chapter 4 of this thesis, we give this new result together with some applications.

1.2.2 The generalized conditional Koziol-Green model under dependent censoring

In the preceding section, we introduced the conditional Koziol-Green estimator, which was developed under two different models. As a first model, the joint distribution function of the lifetime variable and the censoring variable is described by means of an Archimedean copula function. For the second model, the additional information contained in the marginal distribution functions of the lifetime variable and the censoring variable is captured through relation (1.16), which is equivalent to the conditional independence of Z_x and δ_x under dependent censoring. In some applications however, Z_x and δ_x may not be independent, even after conditioning on the covariates. This creates additional challenges in making inference about the marginal conditional distribution function F_x of the lifetime Y_x at a design value $x \in [0,1]$. In the spirit of Section 1.1.1, we overcome this potential difficulty (in the presence of covariates) in the current section by allowing for possible dependence between Z_x and δ_x. In particular, we introduce a

generalization of the conditional Koziol-Green model (1.16) where we assume that the survival function of the censoring variable is a general function of that of the lifetime variable and is given by

$$\bar{G}_x(t) \quad = \quad \mu_x(\bar{F}_x(t)) \ , \ t > 0 \tag{1.18}$$

with $\mu_x(\omega)$ a non-decreasing function of $\omega \in [0,1]$, $\mu_x(0) = 0$ and $\mu_x(1) = 1$. We select this function $\mu_x(\cdot)$ such that the sub-distribution of the uncensored observations satisfies

$$H_x^u(t) \quad = \quad P(Z_x \leq t, \delta_x = 1) = \mathscr{C}_x(\gamma_x, H_x(t)) \tag{1.19}$$

where $\mathscr{C}_x(\cdot,\cdot)$ is some known copula function with the preceding relation following directly from Sklar's theorem, Nelsen (2006). In representation (1.19), $\gamma_x = P(\delta_x = 1)$ is as before, the conditional expected proportion of the uncensored observations at a fixed covariate value $x \in [0,1]$; and $H_x(t) = P(Z_x \leq t)$ is the conditional distribution of the observable time Z_x at $x \in [0,1]$.

To find the function $\mu_x(\cdot)$ we proceed as follows. First, we note from the derivation of the conditional Koziol-Green estimator that

$$H_x^u(t) \quad = \quad \int_0^t \frac{\varphi_x'(\bar{F}_x(s))}{\varphi_x'(H_x(s))} dF(s) = \mathscr{C}_x(\gamma_x, 1 - \bar{H}_x(t)) \tag{1.20}$$

Second, we recall that

$$\bar{H}_x(t) \quad = \quad \varphi_x^{[-1]}\left(\varphi_x(\bar{F}_x(t)) + \varphi_x(\bar{G}_x(t))\right).$$

Substituting the generalized Koziol-Green model (1.18), we obtain

$$\bar{H}_x(t) \quad = \quad \varphi_x^{[-1]}\left(\varphi_x(\bar{F}_x(t)) + \varphi_x(\mu_x(\bar{F}_x(t))\right),$$

which upon plugging into (1.20) yields

$$\begin{aligned} &\int_0^t \frac{\varphi_x'(\bar{F}_x(s))}{\varphi_x'\left(\varphi_x^{[-1]}\left(\varphi_x(\bar{F}_x(s)) + \varphi_x(\mu_x(\bar{F}_x(s)))\right)\right)} dF(s) \\ &\qquad = \quad \mathscr{C}_x\left(\gamma_x, 1 - \varphi_x^{[-1]}\left(\varphi_x(\bar{F}_x(t)) + \varphi_x(\mu_x(\bar{F}_x(t)))\right)\right). \end{aligned} \tag{1.21}$$

Next, we set $\omega = \bar{F}_x(t)$ and after differentiating the preceding equation on both sides and rearranging, we obtain

$$\mu_x'(\omega) \quad = \quad \frac{\varphi_x'(\omega)}{\varphi_x'(\mu_x(\omega))}\left(\frac{1}{\mathscr{C}_{x,01}\left(\gamma_x, 1 - \varphi_x^{[-1]}\left(\varphi_x(\omega) + \varphi_x(\mu_x(\omega))\right)\right)} - 1\right)$$

where $\mu_x'(\omega) = \frac{\partial}{\partial\omega}\mu_x(\omega)$, $\mathscr{C}_{x,ij}(u,v) = \frac{\partial^{i+j}}{\partial u^i \partial v^j}\mathscr{C}_x(u,v)$ denotes the ith and jth partial derivatives of $\mathscr{C}_x(\cdot,\cdot)$ with respect to its first and second arguments respectively, and $\varphi_x'(\cdot)$ is as previously defined.

Furthermore, we define

$$\phi_x(\omega) = \varphi_x^{[-1]}\left(\varphi_x(\omega) + \varphi_x(\mu_x(\omega))\right) \quad \Leftrightarrow \quad \mu_x(\omega) = \varphi_x^{[-1]}\left(\varphi_x(\phi_x(\omega)) - \varphi_x(\omega)\right) \tag{1.22}$$

This implies,

$$\begin{aligned} \phi_x'(\omega) &= \frac{\partial}{\partial\omega}\phi_x(\omega) = \frac{\varphi_x'(\omega) + \varphi_x'(\mu_x(\omega))\mu_x'(\omega)}{\varphi_x'(\phi_x(\omega))} \\ &= \frac{\varphi_x'(\omega)}{\mathscr{C}_{x,01}\left(\gamma_x, 1-\phi_x(\omega)\right)\varphi_x'(\phi_x'(\omega))} \end{aligned}$$

Rearranging and integrating on both sides, we obtain

$$\varphi_x(\omega) = -\int_{\phi_x(\omega)}^{1} \varphi_x'(s)\mathscr{C}_{x,01}\left(\gamma_x, 1-s\right) ds.$$

Subsequently, this leads to

$$\omega = \varphi_x^{[-1]}\left(-\int_{\phi_x(\omega)}^{1} \varphi_x'(s)\mathscr{C}_{x,01}\left(\gamma_x, 1-s\right) ds\right) = \xi_x\left(\gamma_x, \phi_x(\omega)\right).$$

The function $\mu_x(\cdot)$ now follows from (1.22) and is given by

$$\mu_x(\omega) = \varphi_x^{[-1]}\left(\varphi_x\left(\xi_x^{-1}(\gamma_x, \phi_x(\omega))\right) - \varphi_x(\omega)\right),$$

with ξ_x^{-1} denoting the inverse function of $\xi_x\left(\gamma_x, \phi_x(\omega)\right)$ with respect to $\phi_x(\omega)$.

From (1.13), we also note that

$$\begin{aligned} \bar{H}_x(t) &= S_x(t,t) = \varphi_x^{[-1]}\left(\varphi_x(\bar{F}_x(t)) + \mu_x(\bar{F}_x(t))\right) \\ &= \varphi_x^{[-1]}\left(\varphi_x(\bar{F}_x(t)) + \varphi_x\left(\xi_x^{-1}(\gamma_x, \bar{F}_x(t))\right) - \varphi_x(\bar{F}_x(t))\right) \\ &= \xi_x^{-1}(\gamma_x, \bar{F}_x(t)) \end{aligned} \tag{1.23}$$

Consequently, we have

$$\begin{aligned} \bar{F}_x(t) &= \xi_x\left(\gamma_x, \bar{H}_x(t)\right) \\ &= \varphi_x^{[-1]}\left(-\int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}\left(\gamma_x, w\right) dw\right). \end{aligned} \tag{1.24}$$

As before, we now replace γ_x and $H_x(t)$ by their corresponding Stone (1977) type estimators. Subsequently, we obtain the non-parametric survival distribution function estimator

$$\bar{F}_{xh}(t) = \varphi_x^{[-1]}\left(-\int_0^{H_{xh}(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}\left(\gamma_{xh},w\right)dw\right) \tag{1.25}$$

for the lifetime in the generalized conditional Koziol-Green model (1.18) under dependent censoring. It is of interest to point out that the new estimator (1.25) is the conditional version of (1.7) and includes (1.17) as a special case (i.e. when $\mathscr{C}_x(\gamma_{xh},H_{xh}(t)) = \gamma_{xh}H_{xh}(t)$). The nonparametric generalized conditional Koziol-Green estimator (1.25) will be the foundation of Chapter 5. There, we establish some asymptotic (i.e. $n\to\infty$) properties and provide simulations to help get some insight into its finite sample performance.

Before delving into the details of the newly introduced estimators in the consecutive chapters, it is important to emphasize that the Koziol and Green (1976) characterization is a model for informatively censored lifetime data in the sense that it allows the survival distribution of the lifetime to depend on unknown parameters of the survival distribution of the censoring time. This informativeness can be captured indirectly through a relation on the observable time Z and censoring indicator δ. In Table 1.1, we summarize the basic differences between the estimators introduced above for informatively censored lifetime data. For the case without covariates, if we assume that Z and δ are independent, then we have the classical Koziol-Green model (1.1) which together with the independence of the lifetime Y and censoring time C yield F_n^{ACL} as given in (1.2). On the other, if we do not assume that Z and δ are independent, then we get the new extension (1.4), which together with an Archimedean copula to accommodate possible dependence between Y and C lead to the new extended estimator $F_n = 1-\bar{F}_n$ (with $\bar{F}_n$ given in (1.7)). Obviously, we see that F_n is general and includes F_n^{ACL} as a special case. By the same scrutiny for the case with covariates, we also note that $F_{xh} = 1-\bar{F}_{xh}$ (with $\bar{F}_{xh}$ given in (1.25)) is more general and includes the others as special cases.

Table 1.1: Basic interrelations and differences between some estimators under various assumptions on the lifetime Y and censoring time C, in addition to the assumptions on the observable time Z and censoring indicator δ.

Key Assumptions	Estimators	
	Without Covariates	With Covariates
Y and C are independent; Z and δ are independent	F_n^{ACL} – Abdushukurov (1987) and Cheng and Lin (1987)	F_{xh}^{VC} – Veraverbeke and Cadarso-Suárez (2000)
Y and C are dependent; Z and δ are independent	—	F_{xh}^{BV} – Braekers and Veraverbeke (2008)
Y and C are dependent; Z and δ are not independent	F_n – A new proposition	F_{xh} – A new proposition

1.3 Some practical data examples

In this section, we present 3 data sets that will be used to illustrate the practical application of the analysis techniques on which this thesis is based. The first data set is the result of the Worcester Heart Attack Study (WHAS), which has the objective to describe trends over time in the incidence and survival rate following hospital admission of acute myocardial infarction (AMI) patients. The data is collected during ten 1-year periods beginning in 1975 on all AMI patients admitted to hospitals in the Worcester, Massachusetts, metropolitan area. It has information on 8 000 admissions. However, the version of the WHAS data set we utilize in this thesis is taken from the book by Hosmer and Lemeshow (1999). It is a 10% random sample of the original WHAS data set. In this sample, only a small subset of variables is included. Some of these variables are the hospital admission date, the discharge date and the date of last follow-up, from which various survival time variables can be created. Two times that are calculated from these dates and are included in the data set are the length of stay (hospital admission to discharge) and total follow-up (hospital admission to last follow-up). Each has its own

censoring variable that indicates whether the study unit was alive at hospital discharge or last follow-up respectively. Also included in this data set are some key variables which are believed to influence the survival time variables. In addition, subjects with any missing values are dropped from the sampled data set. As a result, the WHAS data we use in this thesis has information on only 481 study units.

The second data set is obtained in a study on size regulation of Atlantic halibuts in the Atlantic coast of Canada following a drastic reduction in the population of the fish species. It aimed at a minimum size limit for retained halibuts for the bottom trawl and long line fishery. However, a minimum size limit would be effective only if an acceptable proportion of the fish returned to the water survive capture, handling and release. For this purpose, the research vessel installed special holding tanks aboard in which the investigators placed the fishes. The measured response for each fish was the time elapsed in hours between placing the fish in the holding tank and death. A remarkable feature about this study is that some animals are removed from the experiment before they die and their faith is unknown. Also, limited holding facilities on board the research vessel necessitated the occasional removal of live experimental animals after 48 hours from the tank for disposal or release in order to accommodate more experimental animals. In addition, all fish surviving past 50-day duration for the experiment were assigned maximum survival time of 1200 hours and treated as right censored observations. In addition to the response variable, the researchers also recorded, for each fish, the covariate fork length of the fish, handling time, total catch weight and depth trawled, which they believed to have an influence on the survival of the fish. For further details about the Atlantic halibut data set, we refer to Lange et al. (1994).

The last data set is the result of a prospective historical clinical study that took place in the period 1962-77 at the University Hospital of Odense, Denmark. It has information on 225 malignant melanoma (cancer of the skin) patients who underwent a surgical operation in which the tumor was completely removed together with the skin within a distance of 2.5 cm around it. In this study, the response variable of interest is the time until malignant melanoma related death. As a result, those patients who did not die of skin cancer were right censored at the study duration or time of death from other causes. Some covatiates recorded at the time of the operation were the sex and age

of the patients. In addition, tumor characteristics such as width, location on the body, thickness, growth patterns, types of malignant cells and ulceration were documented. See Andersen et al. (1993) for a more elaborate description of the melanoma data set.

2

The extended Koziol-Green model under dependent censoring

In Section 1.1.1 of the previous chapter, we introduced the estimator proposed by Abdushukurov (1987) and Cheng and Lin (1987) for informatively censored survival data. This estimator was developed under the classical Koziol and Green (1976) model of random censorship (1.1) and is known to be more efficient than the Kaplan and Meier (1958) estimator. By Csörgó (1988), it was clear that the applicability of the estimator could be limited in practice. To ameliorate the shortfall, we considered an extension of the estimator of Abdushukurov (1987) and Cheng and Lin (1987), in Section 1.1.1. More precisely, we derived a non-parametric estimator for the distribution function of a

survival time under an extension of the classical Koziol-Green model. In the extended Koziol-Green model, we expressed the marginal distribution of the censoring time as a function of the marginal distribution of the survival time, where this function was found through some known copula function on the observable lifetime and the censoring indicator. In order to further increase the scope of applicability of the extended estimator, we additionally allow the censoring time to depends on the survival time through the expression of their joint distribution by an Archimedean copula function.

In this chapter, we further study the extended estimator and establish some of its important asymptotic properties. We summarize these properties as three main theorems in Sections 2.1, 2.2 and 2.3. Before we proceed to these sections, we give the following basic definitions and regularity assumptions that are important in establishing the main results of the chapter.

For the distribution function H, we denote the right end point of its support by $T_H = \inf\{t : H(t) = 1\}$.

(A1) For a copula function $\mathscr{C}(\cdot,\cdot)$, we let $\mathscr{C}_{ij}(u,v) = \frac{\partial^{i+j}}{\partial u^i \partial v^j}\mathscr{C}(u,v)$ denote the ith and jth partial derivatives with respect to its first and second coordinates respectively. For a fixed $\gamma \in (0,1)$, we further assume that $\mathscr{C}_{20}(\gamma,v)$, $\mathscr{C}_{02}(\gamma,v)$ and $\mathscr{C}_{11}(\gamma,v)$ exist and are continuous for all $v \in [0,1]$.

(A2) For the generator of an Archimedean copula φ, we define $\varphi'(u) = \frac{d}{du}\varphi(u)$, $\varphi''(u) = \frac{d^2}{du^2}\varphi(u)$, $\varphi'''(u) = \frac{d^3}{du^3}\varphi(u)$ and assume that $\varphi'''(u)$ exists and is continuous for all $u \in (0,1]$.

In Assumption (A1), γ is the probability of uncensored observations. If $\gamma = 0$, then we have only censored observations and it is not feasible to make inference about the survival distribution of the lifetime. Conversely, $\gamma = 1$ corresponds to fully observable lifetimes, in which case we do not have to account for censoring. Thus, we assume throughout the thesis that $\gamma \in (0,1)$. With this restriction, we see that Assumption (A1) is satisfied by most copula functions. If we take the Gumbel bivariate logistic copula for

example, then for a fixed $\gamma \in (0,1)$, we easily see that

$$\mathscr{C}_{02}(\gamma,v) = -\frac{2\gamma^2(1-\gamma)}{(\gamma+v-\gamma v)^3}, \mathscr{C}_{20}(\gamma,v) = -\frac{2v^2(1-v)}{(\gamma+v-\gamma v)^3} \text{ and } \mathscr{C}_{11}(\gamma,v) = \frac{2\gamma v}{(\gamma+v-\gamma v)^3}$$

exist and are continuous for all $v \in [0,1]$. For the Frank family of copulas, we also find that

$$\begin{aligned}
\mathscr{C}_{02}(\gamma,v) &= -\frac{\theta\left(e^{-\theta\gamma}-1\right)\left(e^{-\theta}-e^{-\theta\gamma}\right)e^{-\theta v}}{\left(\left(e^{-\theta}-1\right)+\left(e^{-\theta\gamma}-1\right)\left(e^{-\theta v}-1\right)\right)^2}, \\
\mathscr{C}_{20}(\gamma,v) &= -\frac{\theta\left(e^{-\theta v}-1\right)\left(e^{-\theta}-e^{-\theta v}\right)e^{-\theta\gamma}}{\left(\left(e^{-\theta}-1\right)+\left(e^{-\theta\gamma}-1\right)\left(e^{-\theta v}-1\right)\right)^2}, \\
\mathscr{C}_{11}(\gamma,v) &= -\frac{\theta e^{-\theta\gamma}e^{-\theta v}\left(e^{-\theta}-1\right)}{\left(\left(e^{-\theta}-1\right)+\left(e^{-\theta\gamma}-1\right)\left(e^{-\theta v}-1\right)\right)^2}
\end{aligned}$$

exist and are continuous for all $v \in [0,1]$, with $\theta \in (-\infty,\infty)$.

Furthermore, we also note that several generators of Archimedean copula functions satisfy Assumption (A2). For the Clayton copula generator for instance

$$\varphi(u) = \frac{1}{\theta}\left(u^{-\theta}-1\right) \quad , \quad \theta \in [-1,\infty)$$

and it is straight forward to see that $\varphi'''(u) = -(1+\theta)(2+\theta)u^{-(3+\theta)}$ exists and is continuous for all $u \in (0,1]$.

2.1 Strong consistency result

The main result of this section is the uniform strong consistency of the extended Koziol-Green estimator $F_n(t)$ as presented in (1.7) in Chapter 1. Also, we obtain the rate of this convergence by means of an exponential bound. These results are summarized as Theorem 2.1 whose proof depends heavily on Lemma 2.1 below. First we give the lemma and afterwards justify its importance in establishing the strong consistency result.

Lemma 2.1. *Suppose* $x \geq 0$, $0 \leq y < 1-H(T)$ *and* $y = \frac{x\varphi'(1)}{2\varphi'(1-H(T)-y)}$. *Then for all* $T < T_H$,

$$\frac{x\varphi'(1)}{2\varphi'\left(\varphi^{[-1]}\left(\varphi(1-H(T))-\frac{x}{2}\varphi'(1)\right)\right)} \leq y \leq 1-H(T)-\varphi^{[-1]}\left(\varphi(1-H(T))-\frac{x}{2}\varphi'(1)\right)$$

Proof. By the mean value theorem, we have

$$\varphi(1-H(T)-y)-\varphi(1-H(T)) \quad = \quad -\varphi'(1-H(T)-y^*)y \tag{2.1}$$

where y^* is a point between zero and y.

Next, we note from the conditions of the Lemma that

$$-\frac{x}{2}\varphi'(1) \quad = \quad -\varphi'(1-H(T)-y)y \quad \geq \quad -\varphi'(1-H(T)-y^*)y. \tag{2.2}$$

Substituting (2.2) into (2.1), gives after some straight forward calculations that

$$y \quad \leq \quad 1-H(T)-\varphi^{[-1]}\left(\varphi(1-H(T))-\frac{x}{2}\varphi'(1)\right). \tag{2.3}$$

Using (2.3), we also get that

$$y \quad = \quad \frac{x\varphi'(1)}{2\varphi'(1-H(T)-y)} \quad \geq \quad \frac{x\varphi'(1)}{2\varphi'\left(\varphi^{[-1]}\left(\varphi(1-H(T))-\frac{x}{2}\varphi'(1)\right)\right)}$$

which concludes the proof. □

Theorem 2.1. *Assume (A1), (A2), $\varphi'(1)<0$ and $T<T_H$, then*

(a) *For all $\varepsilon>0$, we have*

$$P\left(\sup_{0\leq t\leq T}|F_n(t)-F(t)|>\varepsilon\right) \quad \leq \quad 2\exp\left(-\frac{n\alpha^2}{6(3\gamma+\beta)}\right)+D\exp\left(-n\alpha^2\right)$$

where

$$\begin{aligned}\alpha &= \frac{\varphi'(1)\varepsilon}{2\varphi'\left(\varphi^{[-1]}\left(\varphi(1-H(T))-\frac{\varphi'(1)\varepsilon}{2}\right)\right)}\ ,\\ \beta &= 1-H(T)-\varphi^{[-1]}\left(\varphi(1-H(T))-\frac{\varphi'(1)\varepsilon}{2}\right)\end{aligned}$$

and D is a finite positive constant.

(b) *If $n\to\infty$, then*

$$\sup_{0\leq t\leq T}|F_n(t)-F(t)| \quad \to 0 \qquad a.s.$$

Proof. We have that

$$\begin{aligned} F_n(t)-F(t) &= (1-\bar{F}_n(t))-(1-\bar{F}(t)) = \bar{F}(t)-\bar{F}_n(t) \\ &= -\left\{\varphi^{-1}\left(-\int_0^{H_n(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma_n,w)dw\right)\right. \\ &\quad \left.-\varphi^{-1}\left(-\int_0^{H(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma,w)dw\right)\right\}. \end{aligned}$$

Applying the mean value theorem, we get

$$F_n(t)-F(t) = (\gamma_n-\gamma)A(\gamma^*,H^*(t))+(H_n(t)-H(t))B(\gamma^*,H^*(t)),$$

where

$$A(\gamma^*,H^*(t)) = \frac{\int_0^{H^*(t)}\varphi'(1-w)\mathscr{C}_{11}(\gamma^*,w)dw}{\varphi'\left(\varphi^{-1}\left(-\int_0^{H^*(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma^*,w)dw\right)\right)}$$

and

$$B(\gamma^*,H^*(t)) = \frac{\varphi'(1-H^*(t))\mathscr{C}_{01}(\gamma^*,H^*(t))}{\varphi'\left(\varphi^{-1}\left(-\int_0^{H^*(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma^*,w)dw\right)\right)},$$

with γ^* between γ_n and γ, and $H^*(t)$ between $H_n(t)$ and $H(t)$. Using integration by parts, and noting that $\varphi'(1)\mathscr{C}_{10}(\gamma^*,0)=0$, we further obtain

$$A(\gamma^*,H^*(t)) = \frac{\varphi'(1-H^*(t))\mathscr{C}_{10}(\gamma^*,H^*(t))+\int_0^{H^*(t)}\varphi''(1-w)\mathscr{C}_{10}(\gamma^*,w)dw}{\varphi'\left(\varphi^{-1}\left(-\int_0^{H^*(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma^*,w)dw\right)\right)}.$$

Under Assumption $(A1)$, this gives after some calculations that

$$\sup_{0\le t\le T}|A(\gamma^*,H^*(t))| \le \frac{3}{|\varphi'(1)|}\sup_{0\le t\le T}\left|\varphi'(1-H^*(t))\right|. \tag{2.4}$$

Similarly, we also find that

$$\sup_{0\le t\le T}|B(\gamma^*,H^*(t))| \le \frac{1}{|\varphi'(1)|}\sup_{0\le t\le T}\left|\varphi'(1-H^*(t))\right|. \tag{2.5}$$

Using (2.4) and (2.5), we find for all $\varepsilon>0$ and $\eta>0$ that

$$\begin{aligned} P\left(\sup_{0\le t\le T}|F_n(t)-F(t)|>\varepsilon\right) &\le P\left(\frac{3}{|\varphi'(1)|}\sup_{0\le t\le T}\left|\varphi'(1-H^*(t))\right||\gamma_n-\gamma|>\frac{\varepsilon}{2}\right) \\ &\quad +P\left(\frac{1}{|\varphi'(1)|}\sup_{0\le t\le T}\left|\varphi'(1-H^*(t))\right|\sup_{0\le t\le T}|H_n(t)-H(t)|>\frac{\varepsilon}{2}\right), \end{aligned}$$

for which the right hand side of the inequality can be written as

$$P\left(\frac{3}{|\varphi'(1)|}\sup_{0\le t\le T}\left|\varphi'(1-H^*(t))\right||\gamma_n-\gamma|>\frac{\varepsilon}{2},\sup_{0\le t\le T}|H_n(t)-H(t)|\le\eta\right)$$
$$+P\left(\frac{3}{|\varphi'(1)|}\sup_{0\le t\le T}\left|\varphi'(1-H^*(t))\right||\gamma_n-\gamma|>\frac{\varepsilon}{2},\sup_{0\le t\le T}|H_n(t)-H(t)|>\eta\right)$$
$$+P\left(\frac{1}{|\varphi'(1)|}\sup_{0\le t\le T}\left|\varphi'(1-H^*(t))\right|\sup_{0\le t\le T}|H_n(t)-H(t)|>\frac{\varepsilon}{2},\sup_{0\le t\le T}|H_n(t)-H(t)|\le\eta\right)$$
$$+P\left(\frac{1}{|\varphi'(1)|}\sup_{0\le t\le T}\left|\varphi'(1-H^*(t))\right|\sup_{0\le t\le T}|H_n(t)-H(t)|>\frac{\varepsilon}{2},\sup_{0\le t\le T}|H_n(t)-H(t)|>\eta\right).$$

With $0<\eta<1-H(T)$ such that

$$\sup_{0\le t\le T}|\varphi'(1-H^*(t))| \quad < \quad |\varphi'(1-H(T)-\eta)|,$$

the preceding quantity is further bounded above by

$$P\left(|\gamma_n-\gamma|>\frac{\varphi'(1)\varepsilon}{6\varphi'(1-H(T)-\eta)}\right)+P\left(\sup_{0\le t\le T}|H_n(t)-H(t)|>\frac{\varphi'(1)\varepsilon}{2\varphi'(1-H(T)-\eta)}\right)$$
$$+2P\left(\sup_{0\le t\le T}|H_n(t)-H(t)|>\eta\right).$$

Choosing η such that $\eta=\frac{\varphi'(1)\varepsilon}{2\varphi'(1-H(T)-\eta)}$, we easily find that

$$P\left(\sup_{0\le t\le T}|F_n(t)-F(t)|>\varepsilon\right) \quad \le \quad P\left(|\gamma_n-\gamma|>\frac{\eta}{3}\right)+3P\left(\sup_{0\le t\le T}|H_n(t)-H(t)|>\eta\right).$$

Next, we use Bernstein's inequality on the first term at the right hand side of the preceding inequality followed by an application of Dvoretzky, Kiefer and Wolfowitz theorem (see for example, Serfling (1980, page 59)) on the second term of the same inequality. This yields

$$P\left(\sup_{0\le t\le T}|F_n(t)-F(t)|>\varepsilon\right) \quad \le \quad 2\exp\left(-\frac{n\eta^2}{6(3\gamma+\eta)}\right)+D\exp\left(-2n\eta^2\right),$$

where D is a finite positive constant. Using Lemma 2.1, we see that the preceding inequality is further bounded above by

$$2\exp\left(-\frac{n\alpha^2}{6(3\gamma+\beta)}\right)+D\exp\left(-2n\alpha^2\right),$$

with α and β as given in Theorem 2.1.

If we take $\varepsilon_n = \varepsilon = Kn^{-1/2}(\log n)^{1/2}$ for some positive constant K, then it is easy to see that ε_n is small for large n. Thus, by the Borel-Cantelli lemma we find the strong consistency of the extended Koziol-Green estimator. □

2.2 Almost sure asymptotic representation

We now present the extended Koziol-Green estimator as the sum of n independent identically distributed random variable with a remainder term. This representation is an important tool and will pave way for further asymptotic properties of our estimator. It has also been utilized by several authors. For instance, Lo and Singh (1986) employed it for the Kaplan-Meier estimator, Van Keilegom and Veraverbeke (1997) for the Beran estimator, Braekers and Veraverbeke (2008) for the conditional Koziol-Green estimator, among others. Under some conditions, we state such a representation together with the rate of convergence of the remainder term as Theorem 2.2.

Prior to stating the theorem, we give the following lemma which will be used later on. We omit the proof of this lemma, since it is basic and can be found in many standard texts on mathematical statistics. See for example, Serfling (1980).

Lemma 2.2. *Let γ_n and $H_n(t)$ be as previously defined. Then,*

(a) $|\gamma_n - \gamma| = O\left(n^{-1/2}(\log n)^{1/2}\right)$ *a.s.*

(b) $\sup_{0\le t\le T} |H_n(t) - H(t)| = O\left(n^{-1/2}(\log n)^{1/2}\right)$ *a.s.*

Theorem 2.2. *Assume (A1), (A2), $\varphi'(1) < 0$ and $T < T_H$. Then, as $n \to \infty$,*

$$F_n(t) - F(t) = \frac{1}{n}\sum_{i=1}^{n} m_t(Z_i, \delta_i) + r_n(t)$$

where

$$m_t(Z_i,\delta_i) = \frac{1}{\varphi'(F(t))}\left\{(\mathbb{1}\{\delta_i = 1\} - \gamma)\int_0^{H(t)} \varphi'(1-w)\mathscr{C}_{11}(\gamma, w)\,dw \right. \quad (2.6)$$
$$\left. + (\mathbb{1}\{Z_i \le t\} - H(t))\,\varphi'(\bar{H}(t))\mathscr{C}_{01}(\gamma, H(t))\right\}$$

and

$$\sup_{0\le t\le T}|r_n(t)| \;=\; O\left(n^{-1}\log n\right) \quad a.s.$$

Proof. To establish the asymptotic representation of $F_n(t)$, we start with a second order Taylor expansion and obtain

$$\begin{aligned} & F_n(t)-F(t) \\ &= \frac{1}{\varphi'(F(t))}\left\{\int_0^{H_n(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma_n,w)\,dw-\int_0^{H(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma,w)\,dw\right\} \\ &\quad +r_{n1}(t) \end{aligned} \tag{2.7}$$

where

$$\begin{aligned} r_{n1}(t) \;=\; & \frac{\varphi''(\varphi^{-1}(\eta(t)))}{2\left[\varphi'(\varphi^{-1}(\eta(t)))\right]^3}\times \\ & \left\{\int_0^{H_n(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma_n,w)\,dw-\int_0^{H(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma,w)\,dw\right\}^2 \end{aligned}$$

with $\eta(t)$ between $-\int_0^{H_n(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma_n,w)\,dw$ and $-\int_0^{H(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma,w)\,dw$. We denote

$$I(t) \;=\; \int_0^{H_n(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma_n,w)\,dw-\int_0^{H(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma,w)\,dw$$

and find that

$$\sup_{0\le t\le T}|r_{n1}(t)| \;\le\; \frac{1}{|\varphi'(1)|^3}\sup_{0\le t\le T}\varphi''(\varphi^{-1}(\eta(t)))\sup_{0\le t\le T}|I(t)|^2.$$

Using Assumption $(A1)$, it is easy to see that

$$-\int_0^{H_n(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma_n,w)\,dw \quad\text{and}\quad -\int_0^{H(t)}\varphi'(1-w)\mathscr{C}_{01}(\gamma,w)\,dw$$

are respectively bounded above by $\varphi(1-H_n(T))$ and $\varphi(1-H(T))$. On recalling that $\varphi''(\cdot)$ is a decreasing function, we note that the preceding enequality is further bounded above by

$$\frac{1}{|\varphi'(1)|^3}\varphi''(1-H_M(T))\sup_{0\le t\le T}|I(t)|^2$$

where $H_M(T) = \max(H_n(T), H(T))$. Furthermore, we apply the mean value theorem to get

$$I(t) \quad = \quad [\gamma_n - \gamma] \int_0^{H^*(t)} \varphi'(1-w)\mathscr{C}_{11}(\gamma^*, w)dw + [H_n(t) - H(t)]\varphi'(1-H^*(t))\mathscr{C}_{01}(\gamma^*, H^*(t)),$$

with γ^* between γ_n and γ; and $H_n^*(t)$ between $H_n(t)$ and $H(t)$. This gives

$$\begin{aligned} \sup_{0\le t\le T} |I(t)| \quad \le \quad & |\gamma_n - \gamma| \sup_{0\le t\le T} \left| \int_0^{H^*(t)} \varphi'(1-w)\mathscr{C}_{11}(\gamma^*, w)dw \right| \\ & + \sup_{0\le t\le T} |H_n(t) - H(t)| \sup_{0\le t\le T} \left| \varphi'(1-H^*(t))\mathscr{C}_{01}(\gamma^*, H^*(t)) \right|. \end{aligned}$$

Integrating by parts and recalling that $\mathscr{C}_{10}(\gamma^*, 0) = 0$ for all $\gamma^* \in (0,1]$, we obtain

$$\begin{aligned} \sup_{0\le t\le T} \left| \int_0^{H^*(t)} \varphi'(1-w)\mathscr{C}_{11}(\gamma^*, w)dw \right| \quad = \quad & \sup_{0\le t\le T} \left| \varphi'(1-H^*(t))\mathscr{C}_{10}(\gamma^*, H^*(t)) \right| \\ & + \sup_{0\le t\le T} \left| \int_0^{H^*(t)} \varphi''(1-w)\mathscr{C}_{10}(\gamma^*, w)dw \right| \end{aligned}$$

Next, we employ similar deductions as in the proof of Theorem 2.1 and obtain

$$\sup_{0\le t\le T} \left| \int_0^{H^*(t)} \varphi'(1-w)\mathscr{C}_{11}(\gamma^*, w)dw \right| \quad \le \quad 3|\varphi'(1-H_M(T))|.$$

Using the preceding inequality, we get after some calculations that

$$\sup_{0\le t\le T} |I(t)| \quad \le \quad 3|\varphi'(1-H_M(T))||\gamma_n - \gamma| + |\varphi'(1-H_M(T))| \sup_{0\le t\le T} |H_n(t) - H(t)|.$$

Evoking the Glivenko-Cantelli theorem (see Serfling (1980, page 61)), it becomes easy to see that $H_n(T) \to H(T)$ a.s.. Since $H(T) < 1$, we may therefore suppose that $T < T_{H_n}$. In consequent, we obtain by Lemma 2.2 that

$$\sup_{0\le t\le T} |I(t)| \quad = \quad O\left(n^{-1/2}(\log n)^{1/2}\right) \quad \text{a.s.}$$

which subsequently, leads to

$$\sup_{0\le t\le T} |r_{n1}(t)| \quad = \quad O\left(n^{-1}\log n\right) \quad \text{a.s.}$$

We can further decompose the main term in (2.7) by using a second order Taylor expansion to get

$$\begin{aligned}
&\int_0^{H_n(t)} \varphi'(1-w)\mathscr{C}_{01}(\gamma_n,w)\,dw - \int_0^{H(t)} \varphi'(1-w)\mathscr{C}_{01}(\gamma,w)\,dw \\
&\quad = [\gamma_n-\gamma]\int_0^{H(t)} \varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dw + [H_n(t)-H(t)]\,\varphi'(1-H(t))\mathscr{C}_{01}(\gamma,H(t)) \\
&\qquad + r_{n2}(t)+r_{n3}(t)+r_{n4}(t)
\end{aligned} \tag{2.8}$$

where

$$\begin{aligned}
r_{n2}(t) &= \frac{1}{2}[\gamma_n-\gamma]^2\int_0^{H^*(t)} \varphi'(1-w)\mathscr{C}_{21}(\gamma^*,w)\,dw \\
r_{n3}(t) &= \frac{1}{2}[H_n(t)-H(t)]^2\left\{\varphi'(1-H^*(t))\mathscr{C}_{02}(\gamma^*,H^*(t)) - \varphi''(1-H^*(t))\mathscr{C}_{01}(\gamma^*,H^*(t))\right\} \\
r_{n4}(t) &= [\gamma_n-\gamma]\,[H_n(t)-H(t)]\,\varphi'(1-H^*(t))\mathscr{C}_{11}(\gamma^*,H^*(t))
\end{aligned}$$

with γ^* between γ_n and γ; and $H^*(t)$ between $H_n(t)$ and $H(t)$.

Next, we determine the rate of convergence of $r_{n_2}(t), r_{n_3}(t)$ and $r_{n_4}(t)$. Starting with $r_{n2}(t)$, we integrate by parts and obtain

$$\begin{aligned}
r_{n2}(t) &= \frac{1}{2}[\gamma_n-\gamma]^2\Big\{\varphi'(1-H^*(t))\mathscr{C}_{20}(\gamma^*,H_n^*(t)) - \varphi'(1)\mathscr{C}_{20}(\gamma^*,0) \\
&\qquad + \int_0^{H^*(t)} \varphi''(1-w)\mathscr{C}_{20}(\gamma^*,w)dw\Big\}.
\end{aligned}$$

Working as before, we get

$$\sup_{0\le t\le T}|r_{n2}(t)| \le 4|\varphi'(1-H_M(T))|\sup_{0\le v\le 1}|\mathscr{C}_{20}(u,v)|\,[\gamma_n-\gamma]^2 \ , \ \forall u\in(0,1].$$

For $r_{n3}(t)$, we have for all $u\in(0,1]$ that

$$\begin{aligned}
\sup_{0\le t\le T}|r_{n3}(t)| \le &\left\{\varphi'(1-H_M(T))\sup_{0\le v\le 1}|\mathscr{C}_{02}(u,v)| + \varphi''(1-H_M(T))\right\}\times \\
&\sup_{0\le t\le T}|H_n(t)-H(t)|^2.
\end{aligned}$$

Since $H(T)<1$ and $H_n(T)\to H(T)$ a.s. (see Serfling (1980, page 61)), we may suppose again that $T<T_{H_n}$. In consequent, we employ Lemma 2.2 again and obtain

$$\sup_{0\le t\le T}|r_{n2}(t)| = O\left(n^{-1}\log n\right) \text{ a.s. and } \sup_{0\le t\le T}|r_{n3}(t)| = O\left(n^{-1}\log n\right) \text{ a.s.}$$

In the same spirit, we also get that

$$\sup_{0\leq t\leq T}|r_{n4}(t)| \quad = \quad O\left(n^{-1}\log n\right) \qquad \text{a.s.}$$

Next, we let

$$r_n(t) \quad = \quad r_{n1}(t)+\frac{1}{\varphi'(\bar{F}(t))}(r_{n2}(t)+r_{n3}(t)+r_{n4}(t)).$$

From the preceding display, it straight forwardly follows that

$$\sup_{0\leq t\leq T}|r_n(t)| \quad = \quad O\left(n^{-1}\log n\right) \text{ a.s.},$$

since $\varphi'(\bar{F}(t)) \leq \varphi'(1) < 0$ for all $t \in [0,T]$. Using this together with (2.8) and (2.7), concludes the proof. □

2.3 Weak convergence result

As mentioned in the opening of the preceding section, the essence of the almost sure asymptotic representation is to obtain some further asymptotic properties of the extended Koziol-Green estimator. Here, we will establish an additional important property of the estimator. Because of the order of the remainder term given in Theorem 2.2, we will only consider the main term in the asymptotic representation and show that the process associated with the estimator converges weakly to a zero mean Gaussian process with some variance-covariance function, provided $n \to \infty$. We formulate this result as the following theorem.

Theorem 2.3. *Assume the conditions of Theorem 2.2. If $n \to \infty$, then*

$$\sqrt{n}\,(F_n(\cdot)-F(\cdot)) \to W(\cdot)$$

where $W(\cdot)$ is a zero mean Gaussian process with variance-covariance function

$$\begin{aligned}\Gamma(s,t) &= \frac{1}{\varphi'(\bar{F}(s))\varphi'(\bar{F}(t))}\times \\ &\Big\{\gamma(1-\gamma)\int_0^{H(s)}\varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dw\int_0^{H(t)}\varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dw \\ &+(H^u(s)-\gamma H(s))\,\varphi'(\bar{H}(s))\mathscr{C}_{01}(\gamma,H(s))\int_0^{H(t)}\varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dw \\ &+(H^u(t)-\gamma H(t))\,\varphi'(\bar{H}(t))\mathscr{C}_{01}(\gamma,H(t))\int_0^{H(s)}\varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dw \\ &+(H(s\wedge t)-H(s)H(t))\,\varphi'(\bar{H}(s))\varphi'(\bar{H}(t))\mathscr{C}_{01}(\gamma,H(s))\,\mathscr{C}_{01}(\gamma,H(t))\Big\}\end{aligned}$$

Proof. To prove the above theorem, we first establish the finite dimensional distributions of the process $\sqrt{n}(F_n(\cdot)-F(\cdot))$ and then append it with tightness in the space of bounded functions $\ell^\infty[0,T]$. Due to the order of the remainder term in Theorem 2.2 we only have to show the weak convergence of the main term in the asymptotic representation. This translates to showing it for

$$W_n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m_t(Z_i,\delta_i),$$

where $m_t(Z_i,\delta_i)$, $i=1,2,...,n$ are independent copies of $m_t(Z,\delta)$, with $m_t(Z,\delta)$ as defined in (2.6).

After some calculations we get that, for all $t\in[0,T]$, $E[m_t(Z,\delta)]=0$ and, for all $0\le s,t\le T$,

$$\Gamma(s,t) = \mathrm{Cov}[m_s(Z,\delta),m_t(Z,\delta)] = E[m_s(Z,\delta),m_t(Z,\delta)]$$

which equals

$$\begin{aligned}
&\frac{1}{\varphi'(F(s))\varphi'(F(t))}\times\\
&\Big\{\int_0^{H(s)}\varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dw\int_0^{H(t)}\varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dwE\left[(\mathbb{1}\{\delta=1\}-\gamma)^2\right]\\
&+\varphi'(\bar{H}(s))\mathscr{C}_{01}(\gamma,H(s))\int_0^{H(t)}\varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dwE\left[(\mathbb{1}\{\delta=1\}-\gamma)(\mathbb{1}\{Z\le s\}-H(s))\right]\\
&+\varphi'(\bar{H}(t))\mathscr{C}_{01}(\gamma,H(t))\int_0^{H(s)}\varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dwE\left[(\mathbb{1}\{\delta=1\}-\gamma)(\mathbb{1}\{Z\le t\}-H(t))\right]\\
&+\varphi'(\bar{H}(s))\varphi'(\bar{H}(t))\mathscr{C}_{01}(\gamma,H(s))\mathscr{C}_{01}(\gamma,H(t))E\left[(\mathbb{1}\{Z\le s\}-H(s))(\mathbb{1}\{Z\le t\}-H(t))\right]\Big\},
\end{aligned}$$

and gives

$$\begin{aligned}
&\frac{1}{\varphi'(F(s))\varphi'(F(t))}\times\\
&\quad\Big\{\gamma(1-\gamma)\int_0^{H(s)}\varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dw\int_0^{H(t)}\varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dw\\
&\quad+(H^u(s)-\gamma H(s))\,\varphi'(\bar{H}(s))\mathscr{C}_{01}(\gamma,H(s))\int_0^{H(t)}\varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dw\\
&\quad+(H^u(t)-\gamma H(t))\,\varphi'(\bar{H}(t))\mathscr{C}_{01}(\gamma,H(t))\int_0^{H(s)}\varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dw\\
&\quad+(H(s\wedge t)-H(s)H(t))\,\varphi'(\bar{H}(s))\varphi'(\bar{H}(t))\mathscr{C}_{01}(\gamma,H(s))\,\mathscr{C}_{01}(\gamma,H(t))\Big\}.
\end{aligned}$$

Thus, by the multivariate central limit theorem, we get the finite dimensional distributions of the process under consideration.

Next, we show tightness by verifying the conditions of Theorem 2.5.6 of van der Vaart and Wellner (2000). Hereto we show that the class of functions $\mathscr{F}$ given by

$$\mathscr{F}=\{m_t(z,d):t\in[0,T]\}$$

is Donsker.

For the first term in (2.6), we note that $\frac{(\mathbb{1}\{d=1\}-\gamma)}{\varphi'(F(t))}\int_0^{H(t)}\varphi'(1-w)\mathscr{C}_{11}(\gamma,w)\,dw$ is uniformly bounded over t. Furthermore we see that the second function

$$z\to\frac{(\mathbb{1}\{z\le t\}-H(t))}{\varphi'(F(t))}\varphi'(\bar{H}(t))\mathscr{C}_{01}(\gamma,H(t))$$

is uniformly bounded over t and is a monotone function of z, with z and d denoting the observed time and censoring indicator. Hence, we have that $m_t(z,d)$ is a monotone function of z and

$$\begin{aligned}
&\sup_{0\le t\le T} |m_t(z,d)| \\
&= \sup_{0\le t\le T} \left| \frac{(\mathbb{1}\{d=1\}-\gamma)}{\varphi'(F(t))} \left(\varphi'(\bar{H}(t))\mathscr{C}_{10}(\gamma,H(t)) + \int_0^{H(t)} \varphi'(1-w)\mathscr{C}_{10}(\gamma,w)dw \right) \right. \\
&\quad \left. + \frac{(\mathbb{1}\{z\le t\}-H(t))}{\varphi'(F(t))} \varphi'(\bar{H}(t))\mathscr{C}_{01}(\gamma,H(t)) \right| \\
&\le \frac{\varphi'(\bar{H}(T))}{\varphi'(1)} \left(3 + \sup_{0\le t\le T} \mathscr{C}_{01}(\gamma,H(t)) \right) \le M
\end{aligned}$$

where M is a finite positive constant. Using Theorem 2.7.5 of van der Vaart and Wellner (2000), we get that the bracketing number $N_{[]}(\varepsilon,\mathscr{F},L_2(P)) = O\left(\exp\left(\frac{K}{\varepsilon}\right)\right)$ with K a positive constant. Hence, we get that

$$\int_0^\infty \sqrt{\log N_{[]}(\varepsilon,\mathscr{F},L_2(P))}d(\varepsilon) < \infty$$

which shows the class $\mathscr{F}$ is Donsker and completes the proof. □

3

A goodness-of-fit test under the extended Koziol-Green model

In Chapter 1 we introduced the extended Koziol-Green estimator for informatively censored data. It was shown that this estimator is flexible and includes the estimator proposed independently by Abdushukurov (1987) and Cheng and Lin (1987) as a special case. Under some conditions, we established in Chapter 2 that this estimator is uniformly consistent over the sample space. In the same chapter, we found an asymptotic representation for the estimator which lead to the weak convergence of the associated process to a zero mean Gaussian process with some variance-covariance function. In the current chapter, we further pursue the extended Koziol-Green estimator and determine

its validity in practical applications. From the results of Tsiatis (1975), it is apparent that the dependence structure between the censoring time and the survival time cannot be formally investigated, since we do not get to fully observe either of these variables. As a consequence, checking for the validity of the extended Koziol-Green estimator reduces to a goodness-of-fit test for the vertical γ-section of some copula function $\mathscr{C}$ such that characterization (1.4) is satisfied. In light of this, we take as null hypothesis,

$$H_0 : H^u(t) - \mathscr{C}(\gamma, H(t)) = 0 \ , \ \text{for all } t \geq 0 \tag{3.1}$$

versus the general alternative

$$H_a : H^u(t) - \mathscr{C}(\gamma, H(t)) \neq 0, \ \text{for some } t \geq 0$$

In what follows, we provide tools and techniques for the purpose of ascertaining the suitability of the copula function $\mathscr{C}$ under the extended Koziol-Green model. Afterwards, we illustrate the use of the testing procedure on simulated as well as a practical data set.

3.1 Almost sure asymptotic representation

As a basic tool to help obtain the necessary theoretical results for the testing procedure, we first obtain an asymptotic representation of the empirical quantity $H_n^u(t) - \mathscr{C}(\gamma_n, H_n(t))$ as the sum of n independent and identically distributed random variables with a remainder term which is $O\left(n^{-1}\log n\right)$ almost surely. We present this result as Theorem 3.1. Before proceeding, it is important to note that the notations used in this chapter carry directly over from Section 1.1.1 and Chapter 2, unless otherwise stated. Moreover, the results in the present chapter are valid under the regularity assumptions listed in Chapter 2.

Theorem 3.1. *Under the null hypothesis (3.1), assume condition* $(A1)$ *of Chapter 2 is satisfied. Then,* $\forall t \geq 0$

$$H_n^u(t) - \mathscr{C}\left(\gamma_n, H_n(t)\right) = \frac{1}{n}\sum_{i=1}^{n} k_t\left(Z_i, \delta_i\right) + r_n(t)$$

where

$$\begin{aligned} k_t(Z_i,\delta_i) &= \mathbb{1}\{Z_i \le t, \delta_i = 1\} - H^u(t) - (\mathbb{1}\{Z_i \le t\} - H(t))\mathscr{C}_{01}(\gamma, H(t)) \\ &\quad - (\mathbb{1}\{\delta_i = 1\} - \gamma)\mathscr{C}_{10}(\gamma, H(t)) \end{aligned}$$

and

$$\sup_{t\in[0,+\infty]} |r_n(t)| = O\left(n^{-1}\log n\right) \quad a.s.$$

Proof. Under the null hypothesis (3.1), we can write

$$\begin{aligned} H_n^u(t) - \mathscr{C}(\gamma_n, H_n(t)) &= [H_n^u(t) - \mathscr{C}(\gamma_n, H_n(t))] - [H^u(t) - \mathscr{C}(\gamma, H(t))] \\ &= [H_n^u(t) - H^u(t)] - [\mathscr{C}(\gamma_n, H_n(t)) - \mathscr{C}(\gamma, H(t))] \end{aligned}$$

Applying Taylor's expansion on the 2nd term in the preceding equation, we get

$$\begin{aligned} H_n^u(t) - \mathscr{C}(\gamma_n, H_n(t)) &= H_n^u(t) - H^u(t) - [H_n(t) - H(t)]\mathscr{C}_{01}(\gamma, H(t)) \\ &\quad - [\gamma_n - \gamma]\mathscr{C}_{10}(\gamma, H(t)) + r_n(t) \\ &= \frac{1}{n}\sum_{i=1}^{n} k_t(Z_i,\delta_i) + r_n(t) \end{aligned} \tag{3.2}$$

where

$$\begin{aligned} k_t(Z_i,\delta_i) &= \mathbb{1}\{Z_i \le t, \delta_i = 1\} - H^u(t) - [\mathbb{1}\{Z_i \le t\} - H(t)]\mathscr{C}_{01}(\gamma, H(t)) \\ &\quad - [\mathbb{1}\{\delta_i = 1\} - \gamma]\mathscr{C}_{10}(\gamma, H(t)) \end{aligned}$$

and

$$\begin{aligned} r_n(t) &= \frac{1}{2}[\gamma_n - \gamma]^2 \mathscr{C}_{20}(\gamma^*, H^*(t)) + \frac{1}{2}[H_n(t) - H(t)]^2 \mathscr{C}_{02}(\gamma^*, H^*(t)) \\ &\quad + [\gamma_n - \gamma][H_n(t) - H(t)]\mathscr{C}_{11}(\gamma^*, H^*(t)) \end{aligned} \tag{3.3}$$

with γ^* lying between γ_n and γ; and $H^*(t)$ between $H_n(t)$ and $H(t)$. We now determine the rate of convergence of $r_n(t)$. To do so, we note from (3.3) that

$$\begin{aligned} \sup_{t\in[0,+\infty]} |r_n(t)| &\le |\gamma_n - \gamma|^2 \sup_{t\in[0,+\infty]} |\mathscr{C}_{20}(\gamma^*, H^*(t))| \\ &\quad + \sup_{t\in[0,+\infty]} |H_n(t) - H(t)|^2 \sup_{t\in[0,+\infty]} |\mathscr{C}_{02}(\gamma^*, H^*(t))| \\ &\quad + |\gamma_n - \gamma| \sup_{t\in[0,+\infty]} |H_n(t) - H(t)| \sup_{t\in[0,+\infty]} \mathscr{C}_{11}(\gamma^*, H^*(t)) \end{aligned}$$

By Kolmogorov theorem (see for example, Serfling (1980, page 27)) we have that $\gamma_n \to \gamma$ a.s. as $n \to \infty$. From the Glivenko-Cantelli theorem (Serfling (1980, page 61)), we also have $H_n(t) \to H(t)$ a.s as $n \to \infty$. As a result, we know that $\gamma^* \to \gamma$ a.s and $H^*(t) \to H(t)$ a.s as $n \to \infty$. Thus, under Assumption $(A1)$, we can find positive constants M_1, M_2 and M_3 such that

$$\begin{aligned}\sup_{t\in[0,+\infty]} |r_n(t)| &\leq M_1 |\gamma_n - \gamma|^2 + M_2 \sup_{t\in[0,+\infty]} |H_n(t) - H(t)|^2 \\ &\quad + M_3 |\gamma_n - \gamma| \sup_{t\in[0,+\infty]} |H_n(t) - H(t)|\end{aligned}$$

Using Lemma 2.2, it easily follows that

$$\sup_{t\in[0,+\infty]} |r_n(t)| = O\left(n^{-1}\log n\right) \qquad \text{a.s,}$$

which concludes the proof. □

3.2 Weak convergence result

In the previous section, we gave an asymptotic representation of the empirical quantity $H_n^u(\cdot) - \mathscr{C}(\gamma_n, H_n(\cdot))$, since it provides the basis for a valid test of the null hypothesis. As before, we will focus on the main term in the asymptotic representation given in the previous section, under the condition that $n \to \infty$. However, we shall not work with the exact original quantity. Instead, we will work with its normalized version $\sqrt{n}(H_n^u(\cdot) - \mathscr{C}(\gamma_n, H_n(\cdot)))$. In the following theorem, we show the weak convergence of the normalized basic empirical process to a zero mean Gaussian process with a certain variance-covariance function.

Theorem 3.2. *Under the null hypothesis (3.1), suppose Assumption (A1) holds and $n \to \infty$. Then,*

$$\sqrt{n}(H_n^u(\cdot) - \mathscr{C}(\gamma_n, H_n(\cdot))) \to \psi(\cdot) \qquad \textit{in} \qquad \ell^\infty[0,+\infty]$$

where $\psi(\cdot)$ is a zero mean Gaussian process with variance-covariance function given by

$$\begin{aligned}
&\sigma(s,t)\\
&= [H^u(s\wedge t) - H^u(s)H^u(t)] + [H(s\wedge t) - H(s)H(t)]\mathscr{C}_{01}(\gamma,H(s))\mathscr{C}_{01}(\gamma,H(t))\\
&\quad +\gamma[1-\gamma]\mathscr{C}_{10}(\gamma,H(s))\mathscr{C}_{10}(\gamma,H(t)) + [H^u(s) - \gamma H(s)]\mathscr{C}_{01}(\gamma,H(s))\mathscr{C}_{10}(\gamma,H(t))\\
&\quad + [H^u(t) - \gamma H(t)]\mathscr{C}_{01}(\gamma,H(t))\mathscr{C}_{10}(\gamma,H(s)) - [H^u(s\wedge t) - H^u(s)H(t)]\mathscr{C}_{01}(\gamma,H(t))\\
&\quad - [H^u(s\wedge t) - H^u(t)H(s)]\mathscr{C}_{01}(\gamma,H(s)) - [H^u(s) - \gamma H(s)]\mathscr{C}_{10}(\gamma,H(t))\\
&\quad - [H^u(t) - \gamma H(t)]\mathscr{C}_{10}(\gamma,H(s)) \qquad (3.4)
\end{aligned}$$

for all $s \geq 0$ and all $t \geq 0$.

Proof. The proof of Theorem 3.2 proceeds in two steps. First, we establish the convergence of the finite dimensional distributions of the process $\sqrt{n}\,(H_n^u(\cdot) - \mathscr{C}(\gamma_n, H_n(\cdot)))$. Secondly, we show that the process is tight in $\ell^\infty[0,+\infty]$.

To start, we use the main term in the asymptotic representation given in Theorem 3.1 and denote

$$W_n(t) \quad = \quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n} k_t(Z_i,\delta_i).$$

For some integer $q > 0$, we take distinct time points $0 = t_1 < t_2 < \cdots < t_q$. Then, by the multivariate central limit theorem, $(W_n(t_1), W_n(t_2), ..., W_n(t_q))$ converges to an asymptotic normal distribution with mean vector

$$E\,(W_n(t)) \quad = \quad E\,(k_t(Z,\delta)) = 0$$

and variance-covariance matrix equals

$$\begin{aligned}
\sigma(s,t) &= \mathrm{Cov}\,(k_s(Z,\delta)k_t(Z,\delta)) = E\,(k_s(Z,\delta)k_t(Z,\delta))\\
&= [H^u(s\wedge t) - H^u(s)H^u(t)] + [H(s\wedge t) - H(s)H(t)]\mathscr{C}_{01}(\gamma,H(s))\mathscr{C}_{01}(\gamma,H(t))\\
&\quad +\gamma[1-\gamma]\mathscr{C}_{10}(\gamma,H(s))\mathscr{C}_{10}(\gamma,H(t)) + [H^u(s) - \gamma H(s)]\mathscr{C}_{01}(\gamma,H(s))\mathscr{C}_{10}(\gamma,H(t))\\
&\quad + [H^u(t) - \gamma H(t)]\mathscr{C}_{01}(\gamma,H(t))\mathscr{C}_{10}(\gamma,H(s)) - [H^u(s\wedge t) - H^u(s)H(t)]\mathscr{C}_{01}(\gamma,H(t))\\
&\quad - [H^u(s\wedge t) - H^u(t)H(s)]\mathscr{C}_{01}(\gamma,H(s)) - [H^u(s) - \gamma H(s)]\mathscr{C}_{10}(\gamma,H(t))\\
&\quad - [H^u(t) - \gamma H(t)]\mathscr{C}_{10}(\gamma,H(s))
\end{aligned}$$

for all $s = t_j \geq 0$ and $t = t_k \geq 0$.

To show tightness, we first note that

$$\begin{aligned}\sup_{t\in[0,+\infty]}|k_t(Z,\delta)| &\leq \sup_{t\in[0,+\infty]}|\mathbb{1}\{Z\leq t,\delta=1\}-H^u(t)|+|\mathbb{1}\{\delta=1\}-\gamma|\sup_{t\in[0,+\infty]}\mathscr{C}_{10}(\gamma,H(t))\\ &\quad+\sup_{t\in[0,+\infty]}|\mathbb{1}\{Z\leq t\}-H(t)|\sup_{t\in[0,+\infty]}\mathscr{C}_{01}(\gamma,H(t))\\ &\leq 3\end{aligned}$$

Secondly, we define

$$\mathscr{F} = \{k_t(Z,\delta) : t \geq 0\}$$

Then, $\mathscr{F}$ consists of uniformly bounded function over $[0,+\infty]$. As such, their bracketing number is $N_{[]}(\alpha,\mathscr{F},L_2(P))) = O\left(\exp(K\alpha^{-1})\right)$ for $\alpha < 6$ and some $K > 0$. For $\alpha > 6$, we take $N_{[]}(\alpha,\mathscr{F},L_2(P))) = 1$. Furthermore, we note that proving tightness of the process is equivalent to showing that the class of functions $\mathscr{F}$ is Donsker. As a result, we apply Theorem 19.5 of van der Vaart (1998) and obtain

$$\begin{aligned}\int_0^1\sqrt{\log N_{[]}(\alpha,\mathscr{F},L_2(P))}d(3\alpha) &= 3\int_0^1\sqrt{\log N_{[]}(\alpha,\mathscr{F},L_2(P))}d\alpha\\ &\leq 3\int_0^1\sqrt{\frac{K}{\alpha}}d\alpha<\infty\end{aligned}$$

This shows that the process under consideration is tight in $\ell^\infty[0,+\infty]$. Combining this with the convergence of the finite dimensional distributions completes the proof. □

3.3 Goodness-of-fit test statistics

Now, we introduce two goodness of fit test statistics to help investigate the validity of the extended Koziol-Green model in practical applications. Both test statistics are based on the basic empirical process

$$\psi_n(\cdot) = \sqrt{n}(H_n^u(\cdot)-\mathscr{C}(\gamma_n,H_n(\cdot)))$$

More precisely, we consider the Kolmogorov-Smirnov and Cramer-von Mises type statistics, which are respectively defined by

$$T_{KS} = \sup_{t\in[0,+\infty]}|\psi_n(t)| \qquad \text{and} \qquad T_{CM} = \int_0^{+\infty}\psi_n(t)^2 d\mathscr{C}(\gamma_n,H_n(t))$$

As a consequence of Theorem 3.2, we now give the following corollary that will serve as the basis for finding critical values for the test statistics.

Corollary 3.1. *Under the null hypothesis (3.1), assume (A1) holds. Then,*

$$T_{KS} \to \sup_{t\in[0,+\infty]} |\psi(t)|$$
$$T_{CM} \to \int_0^{+\infty} \psi(t)^2 d\mathscr{C}(\gamma, H(t))$$

Proof. To establish the first assertion in the Corollary, we note that

$$\begin{aligned} T_{KS} - \sup_{t\in[0,+\infty]} |\psi(t)| &= \sup_{t\in[0,+\infty]} |\psi_n(t)| - \sup_{t\in[0,+\infty]} |\psi(t)| \\ &\leq \sup_{t\in[0,+\infty]} |\psi_n(t) - \psi(t)| \to 0 \quad \text{a.s} \quad , \quad n \to \infty. \end{aligned}$$

For the second assertion in the corollary, we have

$$\left| T_{CM} - \int_0^{+\infty} \psi(t)^2 d\mathscr{C}(\gamma, H(t)) \right| = \left| \int_0^{+\infty} \psi_n(t)^2 d\mathscr{C}(\gamma_n, H_n(t)) - \int_0^{+\infty} \psi(t)^2 d\mathscr{C}(\gamma, H(t)) \right|$$

Adding and subtracting terms, we get

$$\begin{aligned} \left| T_{CM} - \int_0^{+\infty} \psi(t)^2 d\mathscr{C}(\gamma, H(t)) \right| &\leq \left| \int_0^{+\infty} \left[\psi_n(t)^2 - \psi(t)^2 \right] d\mathscr{C}(\gamma_n, H_n(t)) \right| \\ &\quad + \left| \int_0^{+\infty} \psi(t)^2 d\left[\mathscr{C}(\gamma_n, H_n(t)) - \mathscr{C}(\gamma, H(t)) \right] \right|. \end{aligned} \tag{3.5}$$

But,

$$\left| \int_0^{+\infty} \left[\psi_n(t)^2 - \psi(t)^2 \right] d\mathscr{C}(\gamma_n, H_n(t)) \right| \leq \sup_{t\in[0,+\infty]} \left| \psi_n(t)^2 - \psi^2(t) \right| \int_0^{+\infty} d\mathscr{C}(\gamma_n, H_n(t)).$$

Considering the ordered observed times $z_{(1)}, z_{(2)}, ..., z_{(n)}$, we get that

$$\begin{aligned} \int_0^{+\infty} d\mathscr{C}(\gamma_n, H_n(t)) &= \sum_{r=1}^{n} \left[\mathscr{C}(\gamma_n, \frac{r}{n}) - \mathscr{C}(\gamma_n, \frac{r-1}{n})) \right] \\ &\leq \sum_{r=1}^{n} \left[\frac{r}{n} - \frac{r-1}{n} \right] \leq 1 \end{aligned}$$

where r is the rank of $z_{(r)}$ $(r = 1, 2, ..., n)$. The inequality in the above display follows from Theorem 2.2.4 in Nelsen (2006). Hence, as a consequence of Theorem 3.2, we get

$$\left| \int_0^{+\infty} \left[\psi_n(t)^2 - \psi(t)^2 \right] d\mathscr{C}(\gamma_n, H_n(t)) \right| \to 0 \quad \text{a.s} \quad , \quad n \to \infty \tag{3.6}$$

Next, we recall that as $n \to \infty$, $\gamma_n \xrightarrow{P} \gamma$ a.s and $H_n(t) \xrightarrow{P} H(t)$ a.s. Therefore, using Assumption $(A1)$ together with Lemma 2.2 implies $\mathscr{C}(\gamma_n, H_n(t)) \xrightarrow{P} \mathscr{C}(\gamma, H(t))$ a.s. Further, we note that $\psi(t)^2$ is continuous for all $t \geq 0$. Hence by the Helly-Bray Theorem (Rao (1973, page 117)) we obtain

$$\left| \int_0^{+\infty} \psi(t)^2 d\left[\mathscr{C}(\gamma_n, H_n(t)) - \mathscr{C}(\gamma, H(t))\right] \right| \to 0 \quad \text{a.s} \quad , \quad n \to \infty \tag{3.7}$$

Substituting (3.6) and (3.7) into (3.5) establishes the second assertion in the Corollary. □

For practical application of the test statistics, we propose the following formulation. Let $Z_{(1)}, Z_{(2)}, ..., Z_{(n)}$ denote the order statistics of the Z-sample and $\delta_{(1)}, \delta_{(2)}, ..., \delta_{(n)}$ denote the induced δ-sample. Further, let $r(= 1, 2, ..., n)$ be the rank of $Z_{(r)}$ and denote the number of uncensored observations not greater than $Z_{(r)}$ by

$$N_r \quad = \quad \{1 \leq j \leq r : \delta_{(j)} = 1\}.$$

Then, the test statistics can be expressed as

$$\begin{aligned} T_{KS} &= n^{1/2} \max_{1 \leq r \leq n} \left| \frac{N_r}{n} - \mathscr{C}\left(\gamma_n, \frac{r}{n}\right) \right|, \\ T_{CM} &= n \sum_{r=1}^{n} \left(\frac{N_r}{n} - \mathscr{C}\left(\gamma_n, \frac{r}{n}\right) \right)^2 \left(\mathscr{C}\left(\gamma_n, \frac{r}{n}\right) - \mathscr{C}\left(\gamma_n, \frac{r-1}{n}\right) \right). \end{aligned} \tag{3.8}$$

At this juncture, it is obvious that a valid test of the null hypothesis (3.1) should be based on the null distribution of the test statistics. Due to its complicated variance-covariance function $\sigma(s,t)$ (as given in Theorem 3.2), it is not feasible to readily find critical values for the test. As a resort, we propose a bootstrap approximation to the null distribution of the test statistics. Nonetheless, the validity of the bootstrap can only be assured if the original empirical process $\sqrt{n}(H_n^u(\cdot) - \mathscr{C}(\gamma_n, H(\cdot)))$ and its bootstrap counterpart converge to the same limiting Gaussian process.

3.4 Bootstrap approximation of test statistics

In view of the variance-covariance structure in Theorem 3.2, we now describe a bootstrap procedure to approximate the null distribution of the critical values of the Kolmogorov-

Smirnov and Cramer-von Mises type test statistics given in Section 3.3. Also, we give an asymptotic representation of the bootstrap process. This representation, as in the previous chapter, will aid in establishing the theoretical validity of the bootstrap. Before giving this result, we first describe the bootstrap procedure in the following steps.

1. Given the observed data, we estimate γ and $H(t)$ by

$$\gamma_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\delta_i = 1\} \qquad \text{and} \qquad H_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Z_i \leq t\}$$

2. For each i $(i = 1, 2, ..., n)$,

   (a) we generate two independent uniform (0,1) samples u_i and s_i

   (b) given the copula function $\mathscr{C}$ under the null hypothesis (3.1), we set $v_i = (\mathscr{C}_{10})^{-1}(s_i)$, where $(\mathscr{C}_{10})^{-1}$ is the inverse of $\mathscr{C}_{10}$.

   (c) we define the bootstrap pair (Z_i^*, δ_i^*) by

$$Z_i^* = \inf\{t : H_n(t) \geq v_i\} \qquad \text{and} \qquad \delta_i^* = \mathbb{1}\{u_i > 1 - \gamma_n\}$$

3. We compute the bootstrap counterparts of $H_n^u(t), H_n(t)$ and γ_n respectively by

$$\begin{aligned} H_n^{u^*}(t) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Z_i^* \leq t, \delta_i^* = 1\} \\ H_n^*(t) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Z_i^* \leq t\} \\ \gamma_n^* &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\delta_i^* = 1\} \end{aligned}$$

   and consequently obtain the bootstrap quantity $\left(H_n^{u^*}(t) - \mathscr{C}(\gamma_n^*, H_n^*(t))\right)$.

As previously mentioned, we need to establish the validity of the bootstrap approximation. In light of this, we give an asymptotic representation of the bootstrap empirical process $H_n^{u^*}(\cdot) - \mathscr{C}(\gamma_n^*, H_n^*(\cdot))$ in Section 3.4.1. By means of this valuable tool (i.e. the asymptotic representation), we show in Section 3.4.2 that the bootstrap process and the corresponding original process converge to the same limiting Gaussian process.

3.4.1 Almost sure asymptotic representation of the bootstrap process

Here, we give an asymptotic representation of the bootstrap process under consideration as the sum of n conditionally independent random quantities plus a remainder term which is of order $O\left(n^{-1/2}(\log n)^{1/2}\right)$ almost surely. In the remainder of this chapter, we let P^* and E^* denote Probability and Expectation conditionally on the observed data.

Theorem 3.3. *Under the null hypothesis (3.1), assume (A1) is satisfied. Then,* $\forall t \geq 0$

$$H_n^{u^*}(t) - \mathscr{C}(\gamma_n^*, H_n^*(t)) = \frac{1}{n}\sum_{i=1}^{n}\left\{k_t(Z_i^*, \delta_i^*) - E^* k_t(Z_i^*, \delta_i^*)\right\} + r_n^*(t)$$

where

$$\begin{aligned} k_t(Z_i^*, \delta_i^*) &= \mathbb{1}\{Z_i^* \leq t, \delta_i^* = 1\} - H^u(t) - [\mathbb{1}\{Z_i^* \leq t\} - H(t)]\mathscr{C}_{01}(\gamma, H(t)) \\ &\quad - [\mathbb{1}\{\delta_i^* = 1\} - \gamma]\mathscr{C}_{10}(\gamma, H(t)) \end{aligned}$$

and

$$\sup_{t\in[0,+\infty]} |r_n^*(t)| = O\left(n^{-1/2}(\log n)^{1/2}\right) \qquad a.s$$

Proof. Adding and subtracting terms, we get

$$\begin{aligned} &H_n^{u^*}(t) - \mathscr{C}(\gamma_n^*, H_n^*(t)) \\ &\quad = \left[H_n^{u^*}(t) - \mathscr{C}(\gamma_n^*, H_n^*(t))\right] - [H_n^u(t) - \mathscr{C}(\gamma_n, H_n(t))] + [H_n^u(t) - \mathscr{C}(\gamma_n, H_n(t))] \\ &\quad = \left[H_n^{u^*}(t) - H_n^u(t)\right] - [\mathscr{C}(\gamma_n^*, H_n^*(t)) - \mathscr{C}(\gamma_n, H_n(t))] + [H_n^u(t) - \mathscr{C}(\gamma_n, H_n(t))] \\ &\quad = \left[H_n^{u^*}(t) - H^u(t)\right] - [\mathscr{C}(\gamma_n^*, H_n^*(t)) - \mathscr{C}(\gamma, H(t))] - [H_n^u(t) - H^u(t)] \\ &\qquad + [\mathscr{C}(\gamma_n, H_n(t)) - \mathscr{C}(\gamma, H(t))] + [H_n^u(t) - \mathscr{C}(\gamma_n, H_n(t))] \end{aligned}$$

Using Taylor's expansion on the 2nd and 4th term at the right hand side of the preceding display, we obtain

$$\begin{aligned} &H_n^{u^*}(t) - \mathscr{C}(\gamma_n^*, H_n^*(t)) \\ &\quad = \left\{H_n^{u^*}(t) - H^u(t) - [H_n^*(t) - H(t)]\mathscr{C}_{01}(\gamma, H(t)) - [\gamma_n^* - \gamma]\mathscr{C}_{10}(\gamma, H(t))\right\} \\ &\qquad - \left\{H_n^u(t) - H^u(t) - [H_n(t) - H(t)]\mathscr{C}_{01}(\gamma, H(t)) - [\gamma_n - \gamma]\mathscr{C}_{10}(\gamma, H(t))\right\} \\ &\qquad + R_n^*(t) + R_n(t) \end{aligned} \tag{3.9}$$

where,

$$\begin{aligned} R_n^*(t) &= -\frac{1}{2}\left[H_n^*(t)-H(t)\right]^2 \mathscr{C}_{02}(\gamma_1^\star, H_1^\star(t)) - \frac{1}{2}\left[\gamma_n^*-\gamma\right]^2 \mathscr{C}_{20}(\gamma_1^\star, H_1^\star(t)) \\ &\quad -\left[\gamma_n^*-\gamma\right]\left[H_n^*(t)-H(t)\right]\mathscr{C}_{11}(\gamma_1^\star, H_1^\star(t)) \end{aligned}$$

with $\gamma_1^\star$ lying between γ_n^* and γ, $H_1^\star(t)$ between $H_n^*(t)$ and $H(t)$; and

$$\begin{aligned} R_n(t) &= -\frac{1}{2}\left[H_n(t)-H(t)\right]^2 \mathscr{C}_{02}(\gamma_2^\star, H_2^\star(t)) - \frac{1}{2}\left[\gamma_n-\gamma\right]^2 \mathscr{C}_{20}(\gamma_2^\star, H_2^\star(t)) \\ &\quad -\left[\gamma_n-\gamma\right]\left[H_n(t)-H(t)\right]\mathscr{C}_{11}(\gamma_2^\star, H_2^\star(t)) + \left[H_n^u(t)-\mathscr{C}(\gamma_n, H_n(t))\right] \end{aligned}$$

with $\gamma_2^\star$ between γ_n and γ; and $H_2^\star(t)$ between $H_n(t)$ and $H(t)$. We now give the rate of convergence of $R_n^*(t)$ and $R_n(t)$. Starting with $R_n(t)$, we note that under the null hypothesis (3.1),

$$\begin{aligned} H_n^u(t)-\mathscr{C}(\gamma_n, H_n^u(t)) &= \left[H_n^u(t)-\mathscr{C}(\gamma_n, H_n^u(t))\right] - \left[H^u(t)-\mathscr{C}(\gamma, H(t))\right] \\ &= H_n^u(t)-H^u(t)-\left[\mathscr{C}(\gamma_n, H_n(t))-\mathscr{C}(\gamma, H(t))\right] \end{aligned} \tag{3.10}$$

Applying the bivariate mean value theorem on the 2nd term, we obtain

$$\begin{aligned} H_n^u(t)-\mathscr{C}(\gamma_n, H_n^u(t)) &= \left[H_n^u(t)-H^u(t)\right] - \left[H_n(t)-H(t)\right]\mathscr{C}_{01}(\gamma^*, H^*(t)) \\ &\quad -\left[\gamma_n-\gamma\right]\mathscr{C}_{10}(\gamma^*, H^*(t)) \end{aligned} \tag{3.11}$$

with γ^* lying between γ_n and γ; and $H^*(t)$ between $H_n(t)$ and $H(t)$. Using Dvortzky, Kiefer and Wolfowitz theorem on the first term at the right hand side of (3.11), we have for all $\varepsilon > 0$,

$$P\left(\sup_{t\in[0,+\infty]} |H_n^u(t)-H^u(t)| > \varepsilon\right) \leq C\exp\left(-2n\varepsilon^2\right),$$

with C a finite positive constant. If we take $\varepsilon = \varepsilon_n = Kn^{-1/2}(\log n)^{1/2}$ for some $K > 0$, we get

$$\sum_{n=1}^{\infty} \exp\left(-2n\varepsilon^2\right) < \infty.$$

Thus, by the Borrel-Cantelli lemma, we have

$$\sup_{t\in[0,+\infty]} |H_n^u(t)-H^u(t)| = O\left(n^{-1/2}(\log n)^{1/2}\right) \quad \text{a.s.} \tag{3.12}$$

Next, we have from (3.11) that

$$\begin{aligned}\sup_{t\in[0,+\infty]}|H_n^u(t)-\mathscr{C}(\gamma_n,H_n^u(t))| &\leq \sup_{t\in[0,+\infty]}|H_n^u(t)-H^u(t)|+|\gamma_n-\gamma|\sup_{t\in[0,+\infty]}\mathscr{C}_{10}(\gamma^*,H^*(t))\\ &\quad+\sup_{t\in[0,+\infty]}|H_n(t)-H(t)|\sup_{t\in[0,+\infty]}\mathscr{C}_{01}(\gamma^*,H^*(t))\end{aligned}$$

But, for all $u\in(0,1)$

$$\sup_{0\leq v\leq 1}\mathscr{C}_{10}(u,v) \leq 1 \qquad \text{and} \qquad \sup_{0\leq v\leq 1}\mathscr{C}_{01}(u,v) \leq 1$$

which imply

$$\begin{aligned}&\sup_{t\in[0,+\infty]}|H_n^u(t)-\mathscr{C}(\gamma_n,H_n^u(t))|\\ &\qquad\leq \sup_{t\in[0,+\infty]}|H_n^u(t)-H^u(t)|+\sup_{t\in[0,+\infty]}|H_n(t)-H(t)|+|\gamma_n-\gamma|\end{aligned}$$

Invoking Lemma 2.2 together with (3.12), we obtain

$$\sup_{t\in[0,+\infty]}|H_n^u(t)-\mathscr{C}(\gamma_n,H_n^u(t))| = O\left(n^{-1/2}(\log n)^{1/2}\right) \qquad \text{a.s.}$$

From (3.3), it also follows that

$$\sup_{t\in[0,+\infty]}|R_n(t)| \leq \sup_{t\in[0,+\infty]}|r_n(t)|+\sup_{t\in[0,+\infty]}|H_n^u(t)-\mathscr{C}(\gamma_n,H_n^u(t))|.$$

This gives

$$\begin{aligned}\sup_{t\in[0,+\infty]}|R_n(t)| &= O\left(n^{-1}(\log n)\right)+O\left(n^{-1/2}(\log n)^{1/2}\right) \text{ a.s.}\\ &= O\left(n^{-1/2}(\log n)^{1/2}\right) \text{ a.s.} \qquad (3.13)\end{aligned}$$

For $R_n^*(t)$, we work in analogy with $r_n(t)$ and obtain

$$\sup_{t\in[0,+\infty]}|R_n^*(t)| = O_{P^*}\left(n^{-1}\log n\right) \qquad \text{a.s.} \qquad (3.14)$$

Combining (3.9), (3.12) and (3.13) gives the representation in Theorem 3.3. □

3.4.2 Weak convergence result for the bootstrap process

In this subsection, we show that our bootstrap procedure is a valid process to obtain critical values for the test statistics presented in Section 3.3. As mentioned earlier, this is equivalent to showing that the empirical bootstrap process $\sqrt{n}\left(H_n^{u^*}(\cdot) - \mathscr{C}\left(\gamma_n^*, H_n^*(\cdot)\right)\right)$ and its corresponding original empirical process $\sqrt{n}\left(H_n^{u}(\cdot) - \mathscr{C}\left(\gamma_n, H_n(\cdot)\right)\right)$ converge to the same limiting process. Armed with the almost sure asymptotic representation of the bootstrap process (i.e. Theorem 3.3), we now formulate the validity of the bootstrap procedure in the following theorem.

Theorem 3.4. *Under the null hypothesis H_0, assume (A1) is satisfied. If $n \to \infty$, then*

$$\sqrt{n}\left(H_n^{u^*}(\cdot) - \mathscr{C}\left(\gamma_n^*, H_n^*(\cdot)\right)\right) \to \psi(\cdot) \qquad in \qquad \ell^{\infty}[0,+\infty]$$

where $\psi(\cdot)$ is a zero mean Gaussian process with variance-covariance function $\sigma(s,t)$, given in (3.4).

Proof. Here we work in line with Braekers and Veraverbeke (2005). Let

$$W_n^*(t) \quad = \quad n^{-1/2}\sum_{i=1}^{n}\left\{k_t(Z_i^*, \delta_i^*) - E^* k_t(Z_i^*, \delta_i^*)\right\}.$$

Then, showing the weak convergence of $\sqrt{n}\left(H_n^{u^*}(\cdot) - \mathscr{C}(\gamma_n^*, H_n^*(\cdot))\right)$ is equivalent to that of $W_n^*(\cdot)$, provided $n \to \infty$. To do this, we proceed in two steps. First, we show the convergence of the finite dimensional distributions and later establish tightness of $W_n^*(\cdot)$ in $\ell^{\infty}[0,+\infty]$.

For the convergence of the finite dimension distributions, we show that for any distinct time points $0 < t_1 < t_2 \cdots < t_q$, $q = 1,2,...$

$$\left(W_n^*(t_1), W_n^*(t_2), ..., W_n^*(t_q)\right) \to N\left(0, \sigma(t_j, t_k)\right)$$

Unlike Section 2.3 of Chapter 2, it is not possible to establish the finite dimensional distributions by the multivariate central limit theorem, since the random quantities $k_t(Z_i^*, \delta_i^*)$ $(i = 1,2,...,n)$ are not identically distributed. To this end, we instead verify whether the following two conditions of Araujo and Giné (1980)

1. $\lim_{n\to\infty}\sum_{i=1}^{n} E^*\left(W^*_{n_{ij}}W^*_{n_{ik}}\right) \;=\; \sigma(t_j,t_k)=\sigma_{jk}\;,\;\; 1\le j,k\le q$

2. $\lim_{n\to\infty}\sum_{i=1}^{n}\int_{\left\{|W^*_{n_i}|>\varepsilon\right\}}\left|W^*_{n_i}\right|^2 dP^* \;=\; 0\;,\;\; \forall\varepsilon>0$

hold almost surely for the summands $W^*_{n_{ik}}=n^{-1/2}\left\{k_{t_k}(Z^*_i,\delta^*_i)-E^*k_{t_k}(Z^*_i,\delta^*_i)\right\}$, where

$$\left|W^*_{n_i}\right|^2 \;=\; \sum_{k=1}^{q}\left|W^*_{n_{ik}}\right|^2 \qquad \text{and} \qquad W^*_{n_i} \;=\; \sum_{k=1}^{q} W^*_{n_{ik}}.$$

Adopting similar analogy employed in the calculation of the variance-covariance function of the empirical quantity $\sqrt{n}\,(H^u_n(\cdot)-\mathscr{C}(\gamma_n,H_n(\cdot)))$, we obtain for all $1\le j,k\le q$

$$\begin{aligned}
E^*\left(W^*_{n_{ij}}W^*_{n_{ik}}\right) = \;\frac{1}{n}\Big\{ & [H^u_n(t_j\wedge t_k)-H^u_n(t_j)H^u_n(t_k)] \\
& +[H_n(t_j\wedge t_k)-H_n(t_j)H_n(t_k)]\,\mathscr{C}_{01}(\gamma,H(t_j))\mathscr{C}_{01}(\gamma,H(t_k)) \\
& +\gamma_n[1-\gamma_n]\mathscr{C}_{10}(\gamma,H(t_j))\mathscr{C}_{10}(\gamma,H(t_k)) \\
& +[H^u_n(t_j)-\gamma_n H_n(t_j)]\,\mathscr{C}_{01}(\gamma,H(t_j))\mathscr{C}_{10}(\gamma,H(t_k)) \\
& +[H^u_n(t_k)-\gamma_n H_n(t_k)]\,\mathscr{C}_{01}(\gamma,H(t_k))\mathscr{C}_{10}(\gamma,H(t_j)) \\
& -[H^u(t_j\wedge t_k)-H^u(t_j)H(t_k)]\,\mathscr{C}_{01}(\gamma,H(t_k)) \\
& -[H^u_n(t_j\wedge t_k)-H^u_n(t_k)H_n(t_j)]\,\mathscr{C}_{01}(\gamma,H(t_j)) \\
& -[H^u_n(t_j)-\gamma_n H_n(t_j)]\,\mathscr{C}_{10}(\gamma,H(t_k)) \\
& -[H^u_n(t_k)-\gamma_n H_n(t_k)]\,\mathscr{C}_{10}(\gamma,H(t_j))\Big\}
\end{aligned} \tag{3.15}$$

But, we recall that

$$\begin{aligned}
\sup_{t\in[0,+\infty]}|H^u_n(t)-H^u(t)| \;&=\; O\left(n^{-1/2}(\log n)^{1/2}\right) \qquad \text{a.s.,} \\
\sup_{t\in[0,+\infty]}|H_n(t)-H(t)| \;&=\; O\left(n^{-1/2}(\log n)^{1/2}\right) \qquad \text{a.s.,} \\
|\gamma_n-\gamma| \;&=\; O\left(n^{-1/2}(\log n)^{1/2}\right) \qquad \text{a.s..}
\end{aligned}$$

Assuming $n\to\infty$, we may replace in (3.15) $H^u_n(\cdot),H_n(\cdot)$ and γ_n by $H^u(\cdot),H(\cdot)$ and γ

respectively. Thus almost surely, we get

$$\begin{aligned}
\sigma_{jk} &= \lim_{n\to\infty}\sum_{i=1}^{n} E^*\left(W^*_{n_{ij}}W^*_{n_{ik}}\right)\\
&= [H^u(t_j\wedge t_k) - H^u(t_j)H^u(t_k)] + [H(t_j\wedge t_k) - H(t_j)H(t_k)]\mathscr{C}_{01}(\gamma,H(t_j))\mathscr{C}_{01}(\gamma,H(t_k))\\
&+ \gamma[1-\gamma]\mathscr{C}_{10}(\gamma,H(t_j))\mathscr{C}_{10}(\gamma,H(t_k)) + [H^u(t_j) - \gamma H(t_j)]\mathscr{C}_{01}(\gamma,H(t_j))\mathscr{C}_{10}(\gamma,H(t_k))\\
&+ [H^u(t_k) - \gamma H(t_k)]\mathscr{C}_{01}(\gamma,H(t_k))\mathscr{C}_{10}(\gamma,H(t_j)) - [H^u(t_j\wedge t_k) - H^u(t_j)H(t_k)]\mathscr{C}_{01}(\gamma,H(t_k))\\
&- [H^u(t_j\wedge t_k) - H^u(t_k)H(t_j)]\mathscr{C}_{01}(\gamma,H(t_j)) - [H^u(t_j) - \gamma H(t_j)]\mathscr{C}_{10}(\gamma,H(t_k))\\
&- [H^u(t_k) - \gamma H(t_k)]\mathscr{C}_{10}(\gamma,H(t_j))
\end{aligned}$$

To show that the second condition of Araujo and Giné (1980) holds, we first recall that

$$W^*_{n_{ik}} \;=\; n^{-1/2}\left\{k_{t_k}(Z^*_i,\delta^*_i) - E^*k_{t_k}(Z^*_i,\delta^*_i)\right\} = n^{-1/2}g_{t_k}(Z^*_i,\delta^*_i),$$

where

$$\begin{aligned}
g_{t_k}(Z^*_i,\delta^*_i) \;=\;& \mathbb{1}\{Z^*_i\le t_k,\delta^*_i=1\} - \mathbb{1}\{Z_i\le t_k,\delta_i=1\}\\
&-[\mathbb{1}\{Z^*_i\le t_k\} - \mathbb{1}\{Z_i\le t_k\}]\mathscr{C}_{01}(\gamma,H(t_k))\\
&-[\mathbb{1}\{\delta^*_i=1\} - \mathbb{1}\{\delta_i=1\}]\mathscr{C}_{10}(\gamma,H(t_k)).
\end{aligned}$$

Conditional on the original data, it follows for all $T>0$ that

$$\begin{aligned}
\sup_{0\le t\le T}\left|g_t(Z^*_i,\delta^*_i)\right| \;\le\;& \sup_{0\le t\le T}\left|\mathbb{1}\{Z^*_i\le t,\delta^*_i=1\} - \mathbb{1}\{Z_i\le t,\delta_i=1\}\right|\\
&+ \sup_{0\le t\le T}\left|[\mathbb{1}\{Z^*_i\le t\} - \mathbb{1}\{Z_i\le t\}]\mathscr{C}_{01}(\gamma,H(t))\right|\\
&+ \sup_{0\le t\le T}\left|[\mathbb{1}\{\delta^*_i=1\} - \mathbb{1}\{\delta_i=1\}]\mathscr{C}_{10}(\gamma,H(t))\right|\\
\le\;& 2+\mathscr{C}_{10}(\gamma,H(T)) < \infty
\end{aligned} \tag{3.16}$$

This means that the function $g_{t_k}(Z^*_i,\delta^*_i)$ is uniformly bounded for all $k>0$. Using (3.16), it follows that

$$\begin{aligned}
\max_{1\le i\le n}\left|W^*_{n_i}\right| \;&=\; \max_{1\le i\le n}\left|\sum_{k=1}^{q}W^*_{n_i}\right| = \max_{1\le i\le n}\left|\sum_{k=1}^{q}g_{t_k}(Z^*_i,\delta^*_i)\right|\\
&\le\; n^{-1/2}\max_{1\le i\le n}\sum_{k=1}^{q}\sup_{0\le t\le T}\left|g_t(Z^*_i,\delta^*_i)\right|\\
&=\; n^{-1/2}q\max_{1\le i\le n}\sup_{0\le t\le T}\left|g_t(Z^*_i,\delta^*_i)\right|.
\end{aligned}$$

This implies that

$$\max_{1\leq i\leq n}\left|W_{n_i}^*\right| \quad = \quad O_{P^*}\left(n^{-1/2}\right). \tag{3.17}$$

Similarly, we have

$$\sum_{i=1}^{n}\left|W_{n_i}^*\right|^2 \quad = \quad \sum_{i=1}^{n}\sum_{k=1}^{q}\left|W_{n_{ik}}^*\right|^2 \leq q\max_{1\leq i\leq n}\left(\sup_{0\leq t\leq T}\left|g_t\left(Z_i^*,\delta_i^*\right)\right|\right)^2 = O_{P^*}(1) \tag{3.18}$$

Using (3.17) and (3.18), we get for all $\varepsilon>0$

$$\begin{aligned}\sum_{i=1}^{n}\int_{\left\{\left|W_{n_i}^*\right|>\varepsilon\right\}}\left|W_{n_i}^*\right|^2 dP^* &\leq \int_{\left\{\max_{1\leq i\leq n}\left|W_{n_i}^*\right|>\varepsilon\right\}}\left|W_{n_i}^*\right|^2 dP^* \\ &\leq O_{P^*}(1)P^*\left(\max_{1\leq i\leq n}\left|W_{n_i}^*\right|>\varepsilon\right)=o_{P^*}(1) \qquad \text{a.s}\end{aligned}$$

Hence, the convergence of the finite dimensional distributions.

To prove tightness, we verify the conditions

1. $\sum_{i=1}^{n}E\left[\sup_{t\in\mathscr{F}}|Z_{n_i}(t)|\mathbb{1}\left\{\sup_{t\in\mathscr{F}}|Z_{n_i}(t)|>\lambda\right\}\right]\to 0 \quad , \quad \forall\lambda>0$

2. $\sup_{\rho(t,t')\leq\delta_n}\sum_{i=1}^{n}E\left(Z_{n_i}(t)-Z_{n_i}(t')\right)^2\to 0 \quad , \quad \forall\delta_n\downarrow 0$

3. $\int_0^{\delta_n}\sqrt{\log N_{[]}\left(\varepsilon,\mathscr{F},L_2^n\right)}d\varepsilon\to 0 \quad , \quad \forall\delta_n\downarrow 0$

of the bracketing central limit theorem of van der Vaart and Wellner (2000, Theorem 2.11.9), with $\mathscr{F}$ denoting an index set endowed with an appropriate semimetric ρ.

Let us define

$$X_{n_i}^*(t) \quad = \quad n^{-1/2}k_t(Z_i^*,\delta_i^*) \qquad \text{and} \qquad \mathscr{F} \quad = \quad [0,T].$$

On $\mathscr{F}$, we further define the semimetric

$$\rho\left(t,t'\right) \quad = \quad \max\left(\begin{array}{c}\left|H^u(t)-H^u(t')\right|,\left|H(t)-H(t')\right|,\left|\mathscr{C}_{10}(\gamma,H(t))-\mathscr{C}_{10}(\gamma,H(t'))\right|, \\ \\ \left|\mathscr{C}_{01}(\gamma,H(t))-\mathscr{C}_{01}(\gamma,H(t'))\right|\end{array}\right).$$

Next, we divide $\mathscr{F}$ for every n and ε, conditional on the original observations $Z_1, Z_2, ..., Z_n$ in a partition $\{\mathscr{F}^{*n}_{\varepsilon_j}\}$ such that

$$\sum_{i=1}^{n} E^* \sup_{t,t' \in \mathscr{F}^{*n}_{\varepsilon_j}} \left| X^*_{n_i}(t) - X^*_{n_i}(t') \right|^2 \quad \leq \quad \varepsilon^2 \qquad \text{a.s.} \tag{3.19}$$

The smallest number of intervals of this partition for which (3.19) holds is the bracketing number $N_{[]}(\varepsilon, \mathscr{F}, L_2^n)$.

Given $Z_1, Z_2, ..., Z_n$, we take within $\mathscr{F}$, a sequence of time points $0 = t_0 < t_1 < t_2 < \cdots < t_m$ such that $\{Z_1, Z_2, ..., Z_n\} \subset \{t_0, t_1, ..., t_m\}$ and $\rho(t,t') < C\varepsilon$ for every $t, t' \in [t_{j-1}, t_j]$, $j = 1, 2, ..., m$; where C is a constant to be determine later.

Conditional on the original data, we further define the following partitions

$$\mathscr{F}^{*n}_{\varepsilon_j} \quad = \quad \begin{cases} [t_{j-1}, t_j[& \text{if} \quad t_{j-1} \notin \{Z_1, Z_2, ..., Z_n\} \\ ]t_{j-1}, t_j[& \text{if} \quad t_{j-1} \in \{Z_1, Z_2, ..., Z_n\} \end{cases}$$

$$\mathscr{F}^{*n}_{\varepsilon_m} \quad = \quad \begin{cases} [t_{m-1}, t_j[& \text{if} \quad t_{m-1} \notin \{Z_1, Z_2, ..., Z_n\} \\ ]t_{m-1}, t_j[& \text{if} \quad t_{m-1} \in \{Z_1, Z_2, ..., Z_n\} \end{cases}$$

$$\mathscr{F}^{*n}_{\varepsilon_{j,m+i}} \quad = \quad \{Z_i\} \quad , \quad i = 1, 2, ..., n$$

Next, we show that (3.19) holds for this partition. In this way, the bracketing number $N_{[]}(\varepsilon, \mathscr{F}, L_2^n) = O_{P^*}\left(\frac{1}{\varepsilon}\right) + n$. To do this, we start by noting that

$$\begin{aligned} \left| X^*_{n_i}(t) - X^*_{n_i}(t') \right| \quad &= \quad n^{-1/2} \left| k_t(Z_i^*, \delta_i^*) - k_{t'}(Z_i^*, \delta_i^*) \right| \\ &\leq \quad n^{-1/2} \Big\{ \left| \mathbb{1}\{Z_i^* \leq t, \delta_i^* = 1\} - \mathbb{1}\{Z_i^* \leq t', \delta_i^* = 1\} \right| + \left| H^u(t) - H^u(t') \right| \\ &\qquad + \left| \mathscr{C}_{10}(\gamma, H(t)) - \mathscr{C}_{10}(\gamma, H(t')) \right| + \left| \mathbb{1}\{Z_i^* \leq t\} - \mathbf{1}\{Z_i^* \leq t'\} \right| \\ &\qquad + 2 \left| \mathscr{C}_{01}(\gamma, H(t)) - \mathscr{C}_{01}(\gamma, H(t')) \right| + \left| H(t) - H(t') \right| \Big\}. \end{aligned}$$

Using Cauchy-Schwartz inequality, it follows that

$$\begin{aligned} \left| X^*_{n_i}(t) - X^*_{n_i}(t') \right|^2 \quad &\leq \quad \frac{6}{n} \Big\{ \left| \mathbb{1}\{Z_i^* \leq t, \delta_i^* = 1\} - \mathbb{1}\{Z_i^* \leq t', \delta_i^* = 1\} \right|^2 + \left| H^u(t) - H^u(t') \right|^2 \\ &\qquad + \left| \mathscr{C}_{10}(\gamma, H(t)) - \mathscr{C}_{10}(\gamma, H(t')) \right|^2 + \left| \mathbb{1}\{Z_i^* \leq t\} - \mathbb{1}\{Z_i^* \leq t'\} \right|^2 \\ &\qquad + 4 \left| \mathscr{C}_{01}(\gamma, H(t)) - \mathscr{C}_{01}(\gamma, H(t')) \right|^2 + \left| H(t) - H(t') \right|^2 \Big\}. \end{aligned}$$

Next, we note that

$$\left|\mathbb{1}\{Z_i^* \le t, \delta_i^* = 1\} - \mathbb{1}\{Z_i^* \le t', \delta_i^* = 1\}\right|^2 = \begin{cases} 1 & \text{if} \quad t \wedge t' \le Z_i^* \le t \vee t', \delta_i^* = 1 \\ 0 & \text{otherwise} \end{cases}.$$

This implies that

$$\sup_{t,t' \in \mathscr{F}_{\varepsilon_j}^{*n}} \left|\mathbb{1}\{Z_i^* \le t, \delta_i^* = 1\} - \mathbb{1}\{Z_i^* \le t', \delta_i^* = 1\}\right|^2 = \begin{cases} 1 & \text{if} \quad Z_i^* \in \mathscr{F}_{\varepsilon_j}^{*n} | \{\text{left end point}\} \\ 0 & \text{otherwise} \end{cases}.$$

Also,

$$\left|\mathbb{1}\{Z_i^* \le t\} - \mathbb{1}\{Z_i^* \le t'\}\right|^2 = \begin{cases} 1 & \text{if} \quad t \wedge t' \le Z_i^* \le t \vee t' \\ 0 & \text{otherwise} \end{cases}$$

which gives

$$\sup_{t,t' \in \mathscr{F}_{\varepsilon_j}^{*n}} \left|\mathbb{1}\{Z_i^* \le t\} - \mathbb{1}\{Z_i^* \le t'\}\right|^2 = \begin{cases} 1 & \text{if} \quad Z_i^* \in \mathscr{F}_{\varepsilon_j}^{*n} | \{\text{left end point}\} \\ 0 & \text{otherwise} \end{cases}.$$

Consequently, it follows that

$$E^* \sup_{t,t' \in \mathscr{F}_{\varepsilon_j}^{*n}} \left|\mathbb{1}\{Z_i^* \le t, \delta_i^* = 1\} - \mathbb{1}\{Z_i^* \le t', \delta_i^* = 1\}\right|^2$$
$$= H_n^u\left(\text{right end point of } \mathscr{F}_{\varepsilon_j}^{*n}\right) - H_n^u\left(\text{left end point of } \mathscr{F}_{\varepsilon_j}^{*n}\right) = 0$$

and

$$E^* \sup_{t,t' \in \mathscr{F}_{\varepsilon_j}^{*n}} \left|\mathbb{1}\{Z_i^* \le t\} - \mathbb{1}\{Z_i^* \le t'\}\right|^2$$
$$= H_n\left(\text{right end point of } \mathscr{F}_{\varepsilon_j}^{*n}\right) - H_n\left(\text{left end point of } \mathscr{F}_{\varepsilon_j}^{*n}\right) = 0$$

As pointed out by Braekers and Veraverbeke (2005), this is trivial due to the construction

of the partitions. As a consequence, we have

$$\begin{aligned}
&\sum_{i=1}^{n} E^* \sup_{t,t' \in \mathscr{F}_{\varepsilon_j}^{*n}} \left| X_{n_i}^*(t) - X_{n_i}^*(t') \right|^2 \\
&\qquad \leq \frac{6}{n} \sum_{i=1}^{n} \Big\{ \left| \mathbb{1}\{Z_i^* \leq t, \delta_i^* = 1\} - \mathbb{1}\{Z_i^* \leq t', \delta_i^* = 1\} \right|^2 + \left| H^u(t) - H^u(t') \right|^2 \\
&\qquad\qquad + \left| \mathscr{C}_{10}(\gamma, H(t)) - \mathscr{C}_{10}(\gamma, H(t')) \right|^2 + \left| \mathbb{1}\{Z_i^* \leq t\} - \mathbb{1}\{Z_i^* \leq t'\} \right|^2 \\
&\qquad\qquad + 4 \left| \mathscr{C}_{01}(\gamma, H(t)) - \mathscr{C}_{01}(\gamma, H(t')) \right|^2 + \left| H(t) - H(t') \right|^2 \Big\} \\
&\qquad \leq 42 C^2 \varepsilon^2
\end{aligned}$$

If we take $C^2 = \frac{1}{42}$, we obtain (3.19).

Now we can readily verify the conditions of Theorem 2.11.9 of van der Vaart and Wellner (2000). Starting with the 3rd condition, we take a positive constant D and observe that

$$\int_0^{\zeta_n} \sqrt{\log N_{[]}^*(\varepsilon, \mathscr{F}, L_2^n)} d\varepsilon \leq \int_0^{\zeta_n} \sqrt{\log\left(\frac{D}{\varepsilon} + n\right)} d\varepsilon = \int_0^{\zeta_n} \int_0^{\sqrt{\log\left(\frac{D}{\varepsilon}+n\right)}} d\nu d\varepsilon$$

By Fubini's theorem, this equals $\int_0^\infty f_n(\nu) d\nu$, where

$$f_n(\nu) = \begin{cases} \zeta_n & , \quad \nu \leq \log\left(\frac{D}{\varepsilon} + n\right) \\ \frac{D}{e^{\nu^2} - n} & , \quad \nu > \log\left(\frac{D}{\varepsilon} + n\right) \end{cases}$$

Since f_n converges to zero, we also get pointwise convergence of f_n. Further, we use the bounded convergence theorem (see for example Foran (1991)) and find that, for $\zeta_n \downarrow 0$,

$$\int_0^{\zeta_n} \sqrt{\log N_{[]}^*(\varepsilon, \mathscr{F}, L_2^n)} d\varepsilon \leq \int_0^{\infty} f_n(\nu) d\nu \to 0$$

Also, we have that

$$\begin{aligned}
&\sum_{i=1}^{n} E^* \left(X_{n_i}^*(t) - X_{n_i}^*(t') \right)^2 \\
&\qquad \leq 6 \Big\{ \left| H_n^u(t) - H_n^u(t') \right| + \left| H^u(t) - H^u(t') \right|^2 \\
&\qquad\qquad + \left| \mathscr{C}_{10}(\gamma, H(t)) - \mathscr{C}_{10}(\gamma, H(t')) \right|^2 + \left| H_n(t) - H_n(t') \right| \\
&\qquad\qquad + 4 \left| \mathscr{C}_{01}(\gamma, H(t)) - \mathscr{C}_{01}(\gamma, H(t')) \right|^2 + \left| H(t) - H(t') \right|^2 \Big\}
\end{aligned}$$

So that

$$\sup_{\rho(t,t')<\zeta_n} \sum_{i=1}^{n} E^*\left(X^*_{n_i}(t) - X^*_{n_i}(t')\right)^2 \to 0 \qquad \text{a.s}$$

for every $\zeta_n \downarrow 0$. This gives the 2nd condition of Theorem 2.11.9 of Van der Vaart and Wellner (2000). Before we verify the 3rd condition of the same theorem, we recall that

$$\sup_{t\in\mathscr{F}} \left|X^*_{n_i}(t)\right| \quad \leq \quad n^{-1/2}\left\{2 + \mathscr{C}_{10}(\gamma, H(T))\right\}$$

with $\mathscr{C}_{10}(\gamma, H(T)) \leq 1$ for all $T > 0$. Therefore for all $\eta > 0$, it follows that

$$\mathbb{1}\left\{\sup_{t\in\mathscr{F}} \left|X^*_{n_i}(t)\right| > \eta\right\} \quad = \quad 0$$

if n is sufficiently large. As a result, we get

$$\sum_{i=1}^{n} E^*\left[\sup_{t\in\mathscr{F}} \left|X^*_{n_i}(t)\right| \mathbb{1}\left\{\sup_{t\in\mathscr{F}} \left|X^*_{n_i}(t)\right| > \eta\right\}\right] \to 0 \qquad \text{as} \quad n \to \infty$$

Thus, all the conditions of Theorem 2.11.9 of van der Vaart and Wellner (2000) are satisfied and the quantity $\sqrt{n}\left(H^{u^*}_n(\cdot) - \mathscr{C}(\gamma, H^*_n(\cdot))\right)$ is asymptotically tight. Combining this with the convergence of the finite dimensional distributions concludes the proof. □

3.5 A simulation study

In this section, we set up a simulation study to investigate the finite sample performance of the goodness-of-fit test and its bootstrap approximation of the critical values. Hereto we generate samples of observable couples (Z_i, δ_i), $i = 1, \ldots, n$ such that $H^u(t) = P(Z \leq t, \delta = 1) = \mathscr{C}(\gamma, H(t))$. We assume in this simulation study that the observable lifetimes Z_i $(i = 1, 2, \ldots, n)$ have an exponential distribution ($Z_i \sim \text{Exp}(\lambda)$) with $\lambda = 1.5$ and the indicators δ_i $(i = 1, 2, \ldots, n)$ are Bernoulli distributed with proportion γ of uncensored observations. In particular, we take $\gamma = 25\%, 50\%$ and 75% so as to study the influence of censoring intensity on the bootstrap approximations. Further we use the Clayton copula defined by

$$\mathscr{C}(u,v) \quad = \quad \left[\max\left(u^{-1.3} + v^{-1.3} - 1, 0\right)\right]^{-1/1.3} \quad , \quad \forall (u,v) \in [0,1]^2 \qquad (3.20)$$

to express the relationship between Z_i and δ_i. Using the inverse distribution function method, we obtain our simulation data as follows:

1. We generate two independent uniform $(0,1)$ samples u and t.
2. We set $v = (\mathscr{C}_{10})^{-1}(t)$ where $\mathscr{C}_{10} = \frac{\partial}{\partial u}\mathscr{C}(u,v)$ and $(\mathscr{C}_{10})^{-1}$ is the inverse function of $\mathscr{C}_{10}$
3. We define the observed quantities $d_i = \mathbb{1}\{u > 1-\gamma\}$ and $z_i = -\frac{1}{\lambda}\log(1-v)$.

Based on the simulated data, we utilize the procedure described in Section 4 under the null hypothesis H_0 with the Clayton copula given in (3.20). That is, for a fixed bootstrap size B and for each b $(b = 1,2,\ldots,B)$, we compute $T^*_{KS_b}$ and $T^*_{CM_b}$ based on the bootstrap data $(z^*_1, d^*_1),\ldots,(z^*_n, d^*_n)$, where

$$\begin{aligned} T^*_{KS_b} &= n^{1/2}\max_{1\le r^*\le n}\left|\frac{N_{r^*}}{n} - \mathscr{C}\left(\gamma^*_n, \frac{r^*}{n}\right)\right| \\ T^*_{CM_b} &= n\sum_{r*=1}^{n}\left(\frac{N_{r^*}}{n} - \mathscr{C}\left(\gamma^*_n, \frac{r^*}{n}\right)\right)^2\left(\mathscr{C}\left(\gamma^*_n, \frac{r^*}{n}\right) - \mathscr{C}\left(\gamma^*_n, \frac{r^*-1}{n}\right)\right) \end{aligned} \tag{3.21}$$

with γ^*_n, r^* and N_{r^*} being the bootstrap counterparts of γ_n, r and N_r respectively. Consequently, we obtain an approximate p-value for the test based on T_{KS} and T_{CM} by

$$\frac{1}{B}\sum_{b=1}^{B}\mathbb{1}\left\{T^*_{KS_b} > T_{KS}\right\} \quad \text{and} \quad \frac{1}{B}\sum_{b=1}^{B}\mathbb{1}\left\{T^*_{CM_b} > T_{CM}\right\} \tag{3.22}$$

respectively.

Taking the Product, Plackett and Frank copulas given respectively by

$$\begin{aligned} \mathscr{C}(u,v) &= uv\ , \\ \mathscr{C}(u,v) &= \frac{1}{8}\left\{1+4(u+v) - \sqrt{[1+4(u+v)]^2 - 80uv}\right\}\ , \\ \mathscr{C}(u,v) &= -\frac{1}{4}\log\left(1 + \frac{(e^{-4u}-1)(e^{-4v}-1)}{(e^{-4}-1)}\right)\ , \end{aligned}$$

we also compute the approximate p-values under the corresponding null hypotheses. With a bootstrap size $B = 10000$, we report the P-values of both test statistics in Table 3.1. Since we generated data under (3.20), we expect to conclude the Clayton copula as the most plausible copula function to express the relationship between the observed time and censoring indicator in this simulation study. The results in Table 3.1 show that this is true at 5% level of significance for various degrees of censoring and samples of size 150 or more.

Table 3.1: Approximate p-values (based on 10000 bootstrap replicates) for T_{KS} and T_{CM} under null hypothesis H_0 with Clayton, Product, Plackett and Frank copulas, based on simulated data in which the Clayton copula describes the relationship between the observed time and censoring indicator.

n	γ	Statistic	Clayton	Product	Plackett	Frank
150	25%	T_{KS}	0.4462	0.0000	0.0000	0.0000
		T_{CM}	0.3969	0.0000	0.0000	0.0000
	50%	T_{KS}	0.4603	0.0000	0.0000	0.0040
		T_{CM}	0.3841	0.0000	0.0000	0.0020
	75%	T_{KS}	0.3711	0.0000	0.0013	0.0318
		T_{CM}	0.3292	0.0000	0.0010	0.0182
200	25%	T_{KS}	0.5011	0.0000	0.0000	0.0000
		T_{CM}	0.5134	0.0000	0.0000	0.0000
	50%	T_{KS}	0.4323	0.0000	0.0000	0.0004
		T_{CM}	0.3795	0.0000	0.0000	0.0004
	75%	T_{KS}	0.4326	0.0000	0.0004	0.0110
		T_{CM}	0.3676	0.0000	0.0000	0.0057
250	25%	T_{KS}	0.3526	0.0000	0.0000	0.0000
		T_{CM}	0.2741	0.0000	0.0000	0.0000
	50%	T_{KS}	0.4720	0.0000	0.0000	0.0000
		T_{CM}	0.4175	0.0000	0.0000	0.0000
	75%	T_{KS}	0.3655	0.0000	0.0004	0.0091
		T_{CM}	0.3032	0.0000	0.0000	0.0036

3.6 Data example: Survival with Malignant Melanoma

In this section, we apply the goodness-of-fit test on the melanoma data set, introduced in Chapter 1. The data comes from a historical prospective clinical study conducted in the period 1962-77. The study took place at the university hospital of Odense, Denmark and has information on 225 patients with malignant melanoma (cancer of skin). However, only the 205 patients with complete information are considered here. Of these patients,

57 (28%) died of malignant melanoma (event), 14 (7%) died of other causes and 134 (65%) were alive at the end of the study. See Andersen et al. (1993) for more details about the data set.

For the purpose of this illustration, we treat those observations corresponding to deaths due to other causes and those corresponding to the 134 survivors as censored observations. Before applying the goodness-of-fit test, we perform a preliminary search of a potential copula function $\mathscr{C}$ by graphically investigating whether

$$H_n^u(t) \quad = \quad \mathscr{C}(\gamma_n, H_n(t))$$

nearly holds for all $t \geq 0$, where $H_n^u(t), H_n(t)$ and γ_n are as previously defined. In particular, we compare the vertical γ_n-section of the Fréchet-Hoeffding lower bound (W), Fréchet-Hoeffding upper bound (M), Clayton, Product, Plackett and Frank copulas to the empirical quantity $H_n^u(H_n^{-1}(p))$, where $H_n^{-1}(p) = \inf\{t : H_n(t) > p\}$ is the quantile function of $H_n(t)$. The Clayton, Product, Plackett and Frank copulas are as given in the preceding section and the Fréchet-Hoeffding lower and upper bounds are respectively given by

$$\mathscr{C}(u,v) \quad = \quad \max(u+v-1,0) \text{ and } \mathscr{C}(u,v) \quad = \quad \min(u,v).$$

Table 3.2: Goodness-of-fit test on copula function to describe the relationship between the observed time and censoring indicator in the Malignant Melanoma data set.

	Clayton	Product	Plackett	Frank
T_{KS}	0.3774	1.7682	0.7762	0.6611
P-value	0.2512	0.0000	0.0011	0.0069
T_{CM}	0.0136	0.3355	0.0455	0.0295
P-value	0.0872	0.0000	0.0040	0.0175

Among the copula functions shown in Figure 3.1, we see that the Clayton copula gives the best approximation to the empirical quantity and suggests itself as a potential candidate for this data set. Table 3.2 supports this observation based on the bootstrap procedure described earlier with $B = 10000$ replicates. That is, except for the Clayton copula,

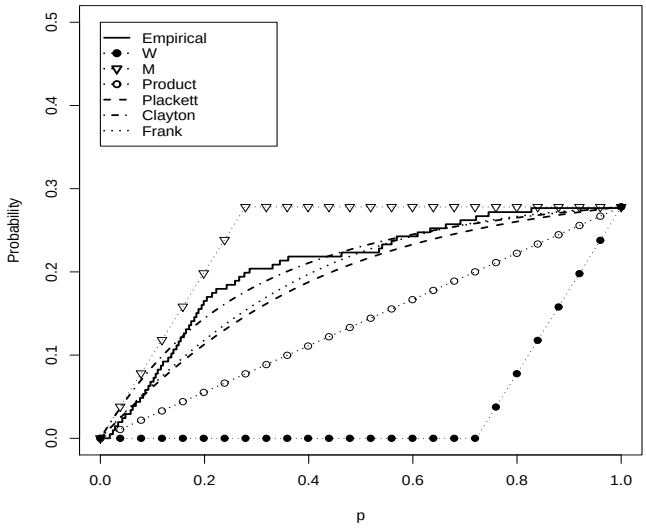

Figure 3.1: Graphical test of the copula function to describe the relationship between the observed time and censoring indicator.

the null hypothesis under the other copula functions is clearly rejected at 5% significance level. This confirms that the Clayton copula function given in Section 5, is appropriate to describe the relationship between the observed survival time and censoring indicator in this data set.

4

The conditional Koziol-Green model under dependent censoring

In Section 1.2.1 of Chapter 1, we introduced the conditional Koziol-Green estimator of Braekers and Veraverbeke (2008). This estimator is a generalization of the conditional Koziol-Green estimator proposed and studied by Veraverbeke and Cadarso-Suárez (2000), where the association between the censoring variable and the lifetime variable is captured by a known Archimedean copula function. In this way, a model which accommodates both dependent and informative censoring was obtained. Braekers and Veraverbeke (2008) derived in this model, a non-parametric Koziol-Green estimator for the conditional distribution function of the lifetime and showed its uniform consistency

and asymptotic normality. In this chapter, we append their results by proving the weak convergence of the process associated with this estimator and give some of its applications. First, we give some regularity conditions in Section 4.1, under which the results of the chapter are valid. In Section 4.2, we give the weak convergence result and present its applications in Section 4.3. We conclude this chapter with a simulation study in Section 4.4 and an illustration of the results on the Worcester heart attack study in Section 4.5.

4.1 Regularity conditions

For the design points $x_1,\ldots,x_n$ we write $\underline{\Delta}_n = \min_{1\le i\le n}(x_i - x_{i-1})$ and $\bar{\Delta}_n = \max_{1\le i\le n}(x_i - x_{i-1})$. The notations $\|K\|_\infty = \sup_{u\in\mathbb{R}} K(u)$, $\|K\|_2^2 = \int_{-\infty}^{+\infty} K^2(u)du$, $\mu_1^K = \int_{-\infty}^{+\infty} uK(u)du$, $\mu_2^K = \int_{-\infty}^{+\infty} u^2K(u)du$ will be used for the kernel K.

We use the following assumptions on the design and on the kernel.

(C1) $x_n \to 1$, $\bar{\Delta}_n = O(n^{-1})$,$\bar{\Delta}_n - \underline{\Delta}_n = o(n^{-1})$.

(C2) K is a probability density function with finite support $[-M,M]$ for some $M > 0$, $\mu_1^K = 0$ and K is Lipschitz of order 1.

The assumption (C1) expresses that the chosen design points are asymptotically equidistant points, selected uniformly over the whole interval $[0,1]$. This implies that, for $c_n(x,h_n)$ defined in Section 1.2, $c_n(x,h_n) = 1$ for n sufficiently large. Therefore we may take $c_n(x,h_n) = 1$ in all proofs of the asymptotic results.

If L is any distribution, then T_L denotes the right endpoint of its support ($T_L = \inf\{t : L(t) = L(+\infty)\}$). We note that $T_{H_x} = T_{F_x} = T_{G_x}$. To obtain our results, we need some smoothness conditions. For a fixed $0 < T < T_{F_x}$,

(C3) $\dot{F}_x(t) = \frac{\partial}{\partial x}F_x(t)$, $\ddot{F}_x(t) = \frac{\partial^2}{\partial x^2}F_x(t)$ exist and are continuous in $(x,t) \in [0,1]\times[0,T]$

(C4) $\dot{\beta}_x = \frac{\partial}{\partial x}\beta_x$, $\ddot{\beta}_x = \frac{\partial^2}{\partial x^2}\beta_x$ exist and are continuous in $x \in [0,1]$

The generator $\varphi_x(v)$ of the Archimedean copula needs to satisfy the following properties.

(C5) $\varphi'_x(v) = \frac{\partial}{\partial v}\varphi_x(v)$ and $\varphi''_x(v) = \frac{\partial^2}{\partial v^2}\varphi_x(v)$ are Lipschitz in the x-direction with a bounded Lipschitz constant, and $\varphi'''_x(v) = \frac{\partial^3}{\partial v^3}\varphi_x(v) \leq 0$ exists and is continuous in $(x,v) \in [0,1]\times]0,1]$.

These assumptions and the fact that φ_x is a generator for an Archimedean copula, give that $\varphi'_x(v)$ is monotone increasing with $\varphi'_x(v) < 0$ and $\varphi''_x(v)$ is monotone decreasing with $\varphi''_x(v) \geq 0$.

4.2 Weak convergence result

In this section, we show the weak convergence of the process $(nh_n)^{1/2}(F^{BV}_{xh}(\cdot) - F_x(\cdot))$ associated with the conditional Koziol-Green estimator $F^{BV}_{xh}(t)$ for the conditional distribution function $F_x(t)$. This adds to the works of Braekers and Veraverbeke (2008), where the authors showed the asymptotic normality in a fixed time point. As in the previous chapters, we first need to derive an almost sure representation for the conditional Koziol-Green estimator $F^{BV}_{xh}(t)$. This result has already been obtained by Braekers and Veraverbeke (2008). For convenience, we formulate their result as the following Lemma.

Lemma 4.1. *Assume conditions (C1)-(C5),* $h_n \to 0$, $\frac{nh_n^5}{\log n} = O(1), T < T_{F_x}$. *Then, for* $t < T_{F_x}$,

$$F^{BV}_{xh}(t) - F_x(t) \quad = \quad \sum_{i=1}^{n} w_{n_i}(x,h_n) m_{tx}(Z_i,\delta_i) + R_n(x,t)$$

where $w_{ni}(x,h_n)$ *is the Gasser-Müller type weight as defined in Section 1.2,*

$$m_{tx}(Z_i,\delta_i) \quad = \quad -\frac{\varphi_x(\bar{H}_x(t))}{\varphi'_x(F_x(t))}(\mathbb{1}\{\delta_i = 1\} - \gamma_x) + \frac{\gamma_x \varphi'_x(\bar{H}_x(t))}{\varphi'_x(F_x(t))}(\mathbb{1}\{Z_i \leq t\} - H_x(t))$$

and as $n \to +\infty$

$$\sup_{0\leq t\leq T} |R_n(x,t)| \quad = \quad O\left((nh_n)^{-1}\log n\right) \qquad a.s.$$

We do not give the prove of the Lemma since it was already established by Braekers and Veraverbeke (2008). Based on the asymptotic representation (i.e. Lemma 4.1), we show the weak convergence of the process $(nh_n)^{1/2}\left(F_{xh}^{BV}(\cdot)-F_x(\cdot)\right)$ in the space $\ell^\infty[0,T]$ of all bounded functions on [0,T] equipped with the supremum-norm. Due to the order of the remainder term in the above representation, we only need to show the weak convergence of the main term in this representation which is the sum of independent quantities of the observed variates. Before we establish the weak convergence result, we give Lemma 4.2 and 4.3 which concern the asymptotic bias and variance respectively.

Lemma 4.2. *Assume (C1), (C2), $F_x(t)$ and β_x satisfy (C3) and (C4) in [0,T] with $T < T_{F_x}$ and φ_x satisfies (C5), $h_n \to 0$. Then, as $n \to +\infty$*

$$\sup_{0\le t\le T}\left|\sum_{i=1}^{n} w_{n_i}(x,h_n)Em_{tx}(Z_i,\delta_i)+\frac{\mu_2^K h_n^2}{2}\left(\frac{\varphi_x(\bar{H}_x(t))}{\varphi_x'(\bar{F}_x(t))}\ddot{\gamma}_x+\frac{\gamma_x\varphi_x'(\bar{H}_x(t))}{\varphi_x'(\bar{F}_x(t))}\ddot{H}_x(t)\right)\right| = o\left(h_n^2\right)+O\left(n^{-1}\right)$$

Proof. For fixed $t \le T$, we have

$$\sum_{i=1}^{n} w_{n_i}(x,h_n)Em_{tx}(Z_i,\delta_i) = -\frac{\varphi_x(\bar{H}_x(t))}{\varphi_x'(F_x(t))}\left(E\gamma_{xh}-\gamma_x\right)+\frac{\gamma_x\varphi_x'(\bar{H}_x(t))}{\varphi_x'(F_x(t))}\left(EH_{xh}(t)-H_x(t)\right)$$

By Lemma A.1.b of Van Keilegom and Veraverbeke (1997a), we get the result. □

Lemma 4.3. *Assume (C1)-(C4) in [0,T] with $T < T_{H_x}$ and φ_x satisfies (C5), $h_n \to 0$, $nh_n \to +\infty$. Then, as $n \to +\infty$*

$$\sup_{0\le t\le T}\left|\sum_{i=1}^{n} w_{ni}^2(x,h_n)Cov\left(m_{sx}(Z_i,\delta_i),m_{tx}(Z_i,\delta_i)\right)-\frac{1}{nh_n}\Gamma_x(s,t)\right| = o\left((nh_n)^{-1}\right)$$

where

$$\Gamma_x(s,t) = ||K||_2^2\left\{\frac{\varphi_x(\bar{H}_x(s))\varphi_x(\bar{H}_x(t))}{\varphi_x'(F_x(s))\varphi_x'(F_x(t))}\gamma_x(1-\gamma_x) + \frac{\gamma_x^2\varphi_x'(\bar{H}_x(s))\varphi_x'(\bar{H}_x(t))}{\varphi_x'(F_x(s))\varphi_x'(F_x(t))}\left(H_x(s\wedge t)-H_x(s)H_x(t)\right)\right\}$$

Proof. From the main term in the asymptotic representation given in Lemma 4.1, we compute for all $0 \leq s,t \leq T$

$$\begin{aligned}
&\mathrm{Cov}\left(m_{sx}(Z_i,\delta_i), m_{tx}(Z_i,\delta_i)\right) \\
&= E\left(m_{sx}(Z_i,\delta_i), m_{tx}(Z_i,\delta_i)\right) - E m_{sx}(Z_i,\delta_i)\, E m_{tx}(Z_i,\delta_i) \\
&= \frac{\varphi_x(\bar{H}_x(s))\varphi_x(\bar{H}_x(t))}{\varphi_x'(F_x(s))\varphi_x'(F_x(t))}\gamma_{x_i}(1-\gamma_{x_i}) + \frac{\gamma_x^2\varphi_x'(\bar{H}_x(s))\varphi_x'(\bar{H}_x(t))}{\varphi_x'(F_x(s))\varphi_x'(F_x(t))}\left(H_{x_i}(s\wedge t) - H_{x_i}(s)H_{x_i}(t)\right) \\
&\quad + \frac{\gamma_x\varphi_x(\bar{H}_x(s))\varphi_x'(\bar{H}_x(t))}{\varphi_x'(F_x(s))\varphi_x'(F_x(t))}\left(H^u_{x_i}(t) - \gamma_{x_i}H_{x_i}(t)\right) + \frac{\gamma_x\varphi_x(\bar{H}_x(t))\varphi_x'(\bar{H}_x(s))}{\varphi_x'(F_x(s))\varphi_x'(F_x(t))}\left(H^u_{x_i}(s) - \gamma_{x_i}H_{x_i}(s)\right)
\end{aligned}$$

By the conditional independent property of Z_i and δ_i, the right hand side of the preceding display reduces to

$$\frac{\varphi_x(\bar{H}_x(s))\varphi_x(\bar{H}_x(t))}{\varphi_x'(F_x(s))\varphi_x'(F_x(t))}\gamma_{x_i}(1-\gamma_{x_i}) + \frac{\gamma_x^2\varphi_x'(\bar{H}_x(s))\varphi_x'(\bar{H}_x(t))}{\varphi_x'(F_x(s))\varphi_x'(F_x(t))}\left(H_{x_i}(s\wedge t) - H_{x_i}(s)H_{x_i}(t)\right)$$

from which the result follows via Lemma 3.1 of Van Keilegom and Veraverbeke (1997a), which is standard in calculating the asymptotic variance function in a fixed design regression setting. □

Theorem 4.1. *Assume conditions (C1)-(C5), $t < T_{F_x}$. Then,*

(a) *If $nh_n^5 \to 0$ and $(nh_n)^{-1/2}\log n \to 0$, then as $n \to +\infty$*

$$(nh_n)^{1/2}\left(F_{xh}(\cdot) - F_x(\cdot)\right) \to W(\cdot|x) \quad \text{in} \quad \ell^\infty[0,T]$$

(b) *If $h_n = Cn^{-1/5}$ for some $C > 0$, then as $n \to +\infty$,*

$$(nh_n)^{1/2}\left(F_{xh}(\cdot) - F_x(\cdot)\right) \to \tilde{W}(\cdot|x) \quad \text{in} \quad \ell^\infty[0,T]$$

where $W(\cdot|x)$ and $\tilde{W}(\cdot|x)$ are Gaussian processes with variance-covariance function $\Gamma_x(s,t)$ as presented in Lemma 4.3, $W(\cdot|x)$ has a zero mean function while for $\tilde{W}$ this is given by

$$b_{tx} = \frac{1}{2}\mu_2^K C^{5/2}\left\{\frac{-\varphi_x(\bar{H}_x(t))}{\varphi_x'(F_x(t))}\ddot{\gamma}_x + \frac{\gamma_x\varphi_x'(\bar{H}_x(t))}{\varphi_x'(F_x(t))}\ddot{H}_x(t)\right\}$$

Proof. From Lemma 4.1 and 4.2, we find

$$F_{xh}(t) - F_x(t) = \sum_{i=1}^n w_{ni}(x,h_n)\xi_{tx}(Z_i,\delta_i) + h_n^2\bar{b}_{tx} + \bar{R}_n(t)$$

where

$$\begin{aligned}\xi_{tx}(Z_i,\delta_i) &= m_{tx}(Z_i,\delta_i) - E\,m_{tx}(Z_i,\delta_i)\\ \sup_{0\le t\le T}|\bar{R}_n(t)| &= O\left((nh_n)^{-3/4}(\log n)^{3/4}\right) + o(h_n^2) \quad \text{a.s.}\end{aligned}$$

and

$$\bar{b}_{tx} = \frac{\mu_2^K h_n^2}{2}\left(\frac{-\varphi_x(\bar{H}_x(t))}{\varphi_x'(\bar{F}_x(t))}\ddot{\gamma}_x + \frac{\gamma_x\varphi_x'(\bar{H}_x(t))}{\varphi_x'(\bar{F}_x(t))}\ddot{H}_x(t)\right).$$

The bias $(nh_n)^{1/2}h_n^2\bar{b}_{tx}$ is $o(1)$ under conditions (a) and equals b_{tx} under conditions (b). Hence it suffices to prove the weak convergence of $W_{hx}(\cdot) = (nh_n)^{1/2}\sum_{i=1}^{n} w_{ni}(x,h_n)\xi_{\cdot x}(Z_i,\delta_i)$ to the Gaussian process $W(\cdot|x)$ with mean zero and covariance function $\Gamma_x(s,t)$.

As before, we do this in two steps. First we show the convergence of the finite dimensional distributions. Next we verify the asymptotic tightness by Theorem 2.11.9 (Bracketing central limit theorem) of van der Vaart and Wellner (2000).

Convergence of the finite dimensional distributions, in this case is that for any $q = 1,2,\ldots$ and any $0 \le t_1 \le \ldots \le t_q \le T$: $(W_{hx}(t_1), W_{hx}(t_2),\ldots,W_{hx}(t_q)) \xrightarrow{D} N(0,\Gamma_x(t_i,t_j))$. Since $W_{hx}(t_i) = \sum_{k=1}^{n} W_{nki}$ where $W_{nki} = (nh_n)^{1/2}w_{nk}(x,h_n)\xi_{t_ix}(Z_k,\delta_k)$, it suffices to check the two conditions of Araujo and Giné (1980) as stated in Section 3.4.2.

Now, applying Lemma 4.3, we obtain

$$\sum_{k=1}^{n} E(W_{nki}W_{nkj}) = (nh_n)\sum_{k=1}^{n} w_{nk}^2(x,h_n)\mathrm{Cov}(m_{t_ix}(Z_k,\delta_k), m_{t_jx}(Z_k,\delta_k)) = \Gamma_x(t_i,t_j) + o(1)$$

Since the functions $\xi_{t_ix}(Z_k,\delta_k)$ are uniformly bounded, it follows that

$$\max_{1\le k\le n}|W_{nk}| = O((nh_n)^{-1/2}) \quad \text{a.s.} \quad \text{and} \quad \sum_{k=1}^{n}|W_{nk}|^2 = O(1) \quad \text{a.s.}$$

Hence,

$$\sum_{k=1}^{n}\int_{\{|W_{nk}|>\varepsilon\}}|W_{nk}|^2 dP \le O(1)P\left(\max_{1\le k\le n}|W_{nk}| > \varepsilon\right) = o(1).$$

To prove the asymptotic tightness, we denote the process $W_{hx}(t)$ as $W_{hx}(t) = \sum_{i=1}^{n} Z_{ni}(t)$ where $Z_{ni}(t) = (nh_n)^{1/2}w_{ni}(x,h_n)\xi_{tx}(Z_i,\delta_i)$.

As before, we need to verify the three conditions of Theorem 2.11.9 of van der Vaart and Wellner (2000). For that purpose, we put on $\mathscr{F} = [0,T]$ and define the semimetric

$$\rho\left(t,t'\right) \quad = \quad \max\left(\begin{array}{c} \left|\frac{-1}{\varphi_x'(F_x(t))} + \frac{1}{\varphi_x'(F_x(t'))}\right|, \;\; |\varphi_x'(\bar{H}_x(t)) - \varphi_x'(\bar{H}_x(t'))|, \\ \\ |\varphi_x(\bar{H}_x(t)) - \varphi_x(\bar{H}_x(t'))|, \; \sup_{x' \in [0,1]} \sqrt{|H_{x'}(t) - H_{x'}(t')|} \end{array}\right)$$

In the third condition, we need the bracketing number $N_{[\,]}(\varepsilon, \mathscr{F}, L_2^n)$. Again, this is the minimal number of sets in a partition of $\mathscr{F} = [0,T] = \bigcup_j \mathscr{F}_{\varepsilon j}$ such that for every set $\mathscr{F}_{\varepsilon j}$,

$$\sum_{i=1}^{n} E\left[\sup_{t,t' \in \mathscr{F}_{\varepsilon j}} |Z_{ni}(t) - Z_{ni}(t')|^2\right] \quad \leq \quad \varepsilon^2.$$

Let us divide $\mathscr{F} = [0,T]$ into subintervals $0 = t_0 \leq t_1 \leq \ldots \leq t_q = T$ where $\rho(t,t') \leq C\varepsilon$ for all $t,t' \in [t_{j-1}, t_j]$, $j = 1, \ldots, q$ with C some constant which we will determine later on. For the partition $\mathscr{F} = [0,t_1] \bigcup \left(\bigcup_{j=2}^{q}]t_{j-1}, t_j]\right)$, we find after some lengthy calculations that

$$\begin{aligned} \left|Z_{ni}(t) - Z_{ni}(t')\right| \quad \leq \quad & (nh_n)^{1/2} w_{ni}(x,h_n) \left(\frac{-1}{\varphi_x'(1)} |\varphi_x(\bar{H}_x(t)) - \varphi_x(\bar{H}_x(t'))| \right. \\ & + (2\varphi_x(\bar{H}_x(T)) + 2\varphi_x'(\bar{H}_x(T))) \left| \frac{-1}{\varphi_x'(F_x(t))} + \frac{1}{\varphi_x'(F_x(t'))} \right| \\ & - \frac{2}{\varphi_x'(1)} |\varphi_x'(\bar{H}_x(t)) - \varphi_x'(\bar{H}_x(t'))| \\ & \left. + \frac{\varphi_x'(\bar{H}_x(T))}{\varphi_x'(1)} \left(|\mathbb{1}\{Z_i \leq t\} - \mathbb{1}\{Z_i \leq t'\}| + |H_{x_i}(t) - H_{x_i}(t')| \right) \right) \end{aligned} \tag{4.1}$$

So

$$\begin{aligned} & \sup_{t,t' \in \mathscr{F}_{\varepsilon j}} \left|Z_{ni}(t) - Z_{ni}(t')\right|^2 \\ & \quad \leq (nh_n) w_{ni}^2(x,h_n) \left\{ C_1 (C\varepsilon)^2 + \left(\frac{\varphi_x'(\bar{H}_x(T))}{\varphi_x'(1)} \right)^2 |\mathbb{1}\{Z_i \leq t_j\} - \mathbb{1}\{Z_i \leq t_{j-1}\}|^2 \right\} \end{aligned}$$

where C_1 is a constant, uniquely determined by the right hand side of (4.1). For the appropriate choice of C, this leads to

$$\sum_{i=1}^{n} E\left[\sup_{t,t'\in\mathscr{F}_{\varepsilon j}} |Z_{ni}(t)-Z_{ni}(t')|^2\right] \quad \leq \quad \varepsilon^2.$$

Hence the bracketing number $N_{[\,]}(\varepsilon,\mathscr{F},L_2^n)$ is equal to $O(\varepsilon^{-1})$ and we get

$$\int_0^{\delta_n} \sqrt{\log N_{[\,]}(\varepsilon,\mathscr{F},L_2^n)}d\varepsilon \quad = \quad \int_0^{\delta_n} \sqrt{\log O(\varepsilon^{-1})}d\varepsilon \to 0$$

when $\delta_n \to 0$. We do not need to verify the second condition of Theorem 2.11.9 of van der Vaart and Wellner (2000), since our partition of $\mathscr{F}=[0,T]$ is independent of n. As last condition we have to check whether for all $\eta>0$,

$$\sum_{i=1}^{n} E\left[\sup_{0\leq t\leq T} |Z_{ni}(t)| \mathbb{1}\left\{\sup_{0\leq t\leq T} |Z_{ni}(t)|>\eta\right\}\right] \to 0 \quad \text{as} \quad n\to+\infty.$$

Since $\xi_{tx}(Z_i,\delta_i)$ is bounded uniformly and $\max\limits_{1\leq i\leq n} w_{ni}(x,h_n)=O((nh_n)^{-1})$ a.s., we get that

$$\sup_{0\leq t\leq T} |Z_{ni}(t)| = O\left((nh_n)^{-1/2}\right) \text{ a.s. },$$

which is always smaller than η for n sufficiently large. So the first condition is also satisfied. Hence, by Theorem 2.11.9 of van der Vaart and Wellner (2000), we have that $W_{hx}(\cdot)\to W(\cdot|x)$ in $\ell^\infty[0,T]$.

□

4.3 Some applications of the weak convergence theorem

The weak convergence result summarized in Theorem 4.1 in the preceding section can be used as a starting point to derive some practical applications. In light of this, we first show in this section that the conditional Koziol-Green estimator is asymptically more efficient in the Koziol-Green model under dependent censoring than the copula-graphic estimator of Braekers and Veraverbeke (2005). A second application is an asymptotic confidence band for the conditional Koziol-Green estimator.

4.3.1 Asymptotic efficiency

At a fixed design point $x \in [0,1]$, Braekers and Veraverbeke (2005) derived the variance-covariance function

$$\begin{aligned}\sigma_x(s,t) = \frac{||K||_2^2}{\varphi_x'(F_x(s))\varphi_x'(F(t))} \Bigg\{ & \int_0^{s\wedge t} \varphi_x'(\bar{H}_x(z))^2 dH_x^u(z) \\ & + \int_0^{s\wedge t} \left(\varphi_x''(\bar{H}_x(w))\bar{H}_x(w) + \varphi_x'(\bar{H}_x(w))\right) \int_0^w \varphi_x''(\bar{H}_x(y)) dH_x^u(y) dH_x^u(w) \\ & + \int_0^{s\wedge t} \varphi_x''(\bar{H}_x(w)) \int_0^{s\wedge t} \left(\varphi_x''(\bar{H}_x(y))\bar{H}_x(y) + \varphi_x'(\bar{H}_x(y))\right) dH_x^u(y) dH_x^u(w) \\ & - \int_0^t \left(\varphi_x(\bar{H}_x(y))\bar{H}_x(y) + \varphi_x'(\bar{H}_x(y))\right) dH_x^u(y) \int_0^s \left(\varphi_x''(\bar{H}(w))\bar{H}_x(w) + \varphi_x'(\bar{H}_x(w))\right) dH_x^u(w) \Bigg\}\end{aligned}$$

At any fixed time point t, it is easy to see that the asymptotic variance of the copula-graphic estimator of Braekers and Veraverbeke (2005) reduces, after some lengthy but straightforward derivation to the expression

$$\begin{aligned}\sigma_x(t,t) = & \frac{||K||_2^2}{\varphi_x'(F(t))^2} \times \\ & \left\{ \gamma_x(1-\gamma_x) \int_0^t \varphi_x'(\bar{H}_x(w))^2 dH_x(w) + \gamma_x^2 \varphi_x'(\bar{H}_x(t))^2 H_x(t)(1-H_x(t)) \right\} \end{aligned} \quad (4.2)$$

when the Koziol-Green model is satisfied.

Analogously, we also obtain the expression

$$\Gamma_x(t,t) = ||K||_2^2 \left\{ \frac{\varphi_x(\bar{H}_x(t))^2}{\varphi_x'(\bar{F}_x(t))^2} \gamma_x(1-\gamma_x) + \frac{\gamma_x^2 \varphi_x'(\bar{H}_x(t))^2}{\varphi_x'(F_x(t))^2} \left(H_x(t)(1-H_x(t))\right) \right\} \quad (4.3)$$

for the asymptotic variance of the conditional Koziol-Green estimator. To show the efficiency of the conditional Koziol-Green estimator over the copula-graphic estimator, we compare expressions (4.2) and (4.3) and get that

$$\begin{aligned}\frac{\Gamma_x(t,t)}{\sigma_x(t,t)} &= \frac{\gamma_x(1-\gamma_x)\varphi_x(\bar{H}_x(t))^2 + \gamma_x^2 \varphi_x'(\bar{H}_x(t))^2 H_x(t)(1-H_x(t))}{\gamma_x(1-\gamma_x)\int_0^t \varphi_x'(\bar{H}_x(s))^2 dH_x(s) + \gamma_x^2 \varphi_x'(\bar{H}_x(t))^2 H_x(t)(1-H_x(t))} \\ &= \frac{\gamma_x(1-\gamma_x)\left(\int_{\bar{H}_x(t)}^1 |\varphi_x'(w)| dw\right)^2 + \gamma_x^2 \varphi_x'(\bar{H}_x(t))^2 H_x(t)(1-H_x(t))}{\gamma_x(1-\gamma_x)\int_0^t \varphi_x'(\bar{H}_x(s))^2 dH_x(s) + \gamma_x^2 \varphi_x'(\bar{H}_x(t))^2 H_x(t)(1-H_x(t))} \\ &\leq \frac{\gamma_x(1-\gamma_x)H_x(t)\int_0^t \varphi_x'(\bar{H}_x(s))^2 dH_x(s) + \gamma_x^2 \varphi_x'(\bar{H}_x(t))^2 H_x(t)(1-H_x(t))}{\gamma_x(1-\gamma_x)\int_0^t \varphi_x'(\bar{H}_x(s))^2 dH_x(s) + \gamma_x^2 \varphi_x'(\bar{H}_x(t))^2 H_x(t)(1-H_x(t))} \leq 1\end{aligned}$$

where the inequality follows from the Cauchy-Schwartz inequality. From this,we note that the upper bound goes to 1 if $\gamma_x \to 1$. This was expected since the estimators in both models become a conditional empirical distribution function when there is no censoring. Also, we see that this upper bound is 1 when $t \to +\infty$ and is $H_x(t)$ when $\gamma_x \to 0$. For a pictorial representation of the relative asymptotic efficiency of the conditional Koziol-Green estimator over the copula-graphic estimator, we present in Figure 4.1, the upper bound for three Archimedean copulas, the independent copula ($\varphi_x(t) = -\log(t)$), the Fréchet-Hoeffding lower bound ($\varphi_x(t) = 1-t$) and the Clayton family copula with $\theta = 1$ ($\varphi_x(t) = \frac{1}{t} - 1$). We use in this picture the conditional distribution function $H_x(t)$ to transform the time-axis to $[0,1]$.

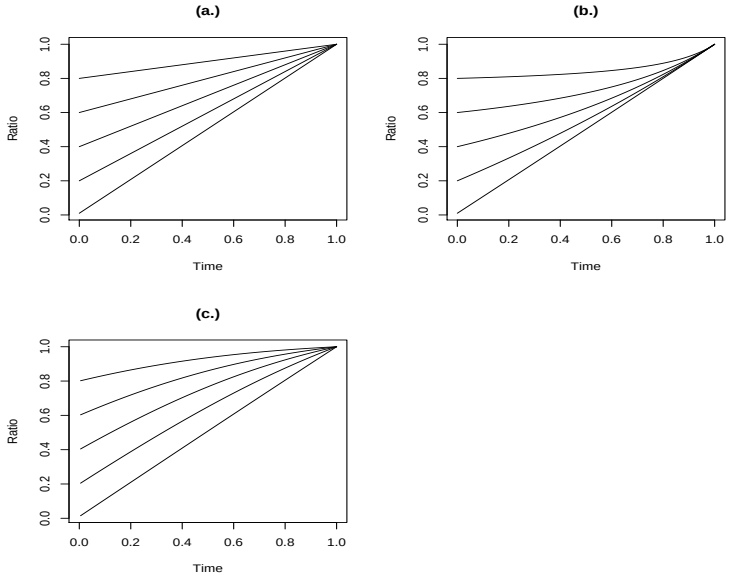

Figure 4.1: The upperbound for the ratio of variances, given for the independent (a.), Fréchet - Hoeffding lower bound (b.) and Clayton family copula ($\theta = 1$) (c.). Each curve presents a different percentage of uncensored observations (bottom till top: $p_{x1} = 0.01, 0.2, 0.4, 0.6, 0.8$).

For the independent copula, we see in Figure 4.1 straight lines for each level of censoring. The Fréchet-Hoeffding lower bound which expresses a discordant association gives convex lines while the concordant Clayton copula shows concave lines. In each plot, we have that the lines converge to 1 at the right end and all curves are lying between the

diagonal and the horizontal line at 1.

4.3.2 Asymptotic confidence band

As a second application of the weak convergence result in Theorem 4.1, we derive an asymptotic confidence band for the conditional Koziol-Green estimator $F_{xh}^{BV}(t)$. Like in the work of Hollander and Pĕna (1989), we introduce an extra parameter λ such that we have a family of bands and which gives some flexibility in the construction of the confidence band. For example, by selecting certain values for λ we can find a more narrow asymptotic confidence band when the sample size is small or moderate, or a more conservative band when we are interested in a time t near the end of the support. We summarize this result as the following theorem.

Theorem 4.2. *Assume the conditions (C1) - (C5) with* $T < T_{F_x}$, $nh_n^5 \to 0$, $(nh_n)^{-1/2}\log n \to 0$ *and* $\lambda > 0$. *For each* $0 < \alpha < 1$, *let* $c_{\alpha xh}$ *be such that, as* $n \to +\infty$,

$$P\left(\sup_{0\le t\le T}\left|B_1(L_{xh}(t)) + \frac{\lambda^{1/2}\varphi_x(\bar{H}_{xh}(t))}{\gamma_{xh}\varphi_x'(\bar{H}_{xh}(t))(H_{xh}(t)+\lambda H_{xh}(t))}B_2(\gamma_{xh})\right| \le c_{\alpha xh}\right) \to 1-\alpha \tag{4.4}$$

Then, as $n \to +\infty$,

$$P(F_{xh}(t) - c_{\alpha xh}D_{xh}(t) \le F_x(t) \le F_{xh}(t) + c_{\alpha xh}D_{xh}(t), \textit{for all } 0 \le t \le T) \to 1-\alpha$$

where $B_1(s)$ *and* $B_2(s)$ *are independent Brownian bridges and*

$$\begin{aligned} L_{xh}(t) &= \frac{\lambda H_{xh}(t)}{H_{xh}(t)+\lambda H_{xh}(t)} \\ D_{xh}(t) &= (nh_n\lambda)^{-1/2}\|K\|_2\frac{\gamma_{xh}\varphi_x'(\bar{H}_{xh}(t))(\bar{H}_{xh}(t)+\lambda H_{xh}(t))}{\varphi_x'(F_{xh}(t))}. \end{aligned}$$

Proof. We note that we can rewrite in Theorem 4.1 the Gaussian process $W(\cdot|x)$ as, for a given $\lambda > 0$,

$$\lambda^{-1/2}\|K\|_2\frac{\gamma_x\varphi_x'(\bar{H}_x(t))(\bar{H}_x(t)+\lambda H_x(t))}{\varphi_x'(F_x(t))}B_1(L_x(t)) + \|K\|_2\frac{\varphi_x(\bar{H}_x(t))}{\varphi_x'(F_x(t))}B_2(\gamma_x)$$

where $\{B_1(s)|0 \le s \le 1\}$ and $\{B_2(s)|0 \le s \le 1\}$ are independent Brownian bridges and

$$L_x(t) = \frac{\lambda H_x(t)}{H_x(t) + \lambda H_x(t)} \tag{4.5}$$

Using Theorem 4.1 together with Theorem 1 of Braekers and Veraverbeke (2008), Lemma A.2 of Van Keilegom and Veraverbeke (1997a), Lemma A.1. of Braekers and Veraverbeke (2001) and Slutsky's Theorem, we have that

$$(F_{xh}(\cdot) - F_x(\cdot))D_{xh}^{-1}(\cdot) \to B_1(L_x(\cdot)) + \frac{\lambda^{1/2}\varphi_x(\bar{H}_x(\cdot))}{\gamma_x \varphi_x'(H_x(\cdot))(H_x(\cdot) + \lambda H_x(\cdot))} B_2(\gamma_x) \quad \text{in} \quad \ell^\infty[0,T].$$

Analogously, we find that

$$\begin{aligned} &B_1(L_{xh}(\cdot)) + \tfrac{\lambda^{1/2}\varphi_x(\bar{H}_{xh}(\cdot))}{\gamma_{xh}\varphi_x'(H_{xh}(\cdot))(H_{xh}(\cdot)+\lambda H_{xh}(\cdot))} B_2(\gamma_{xh}) \\ &\qquad \to B_1(L_x(\cdot)) + \tfrac{\lambda^{1/2}\varphi_x(\bar{H}_x(\cdot))}{\gamma_x\varphi_x'(H_x(\cdot))(H_x(\cdot)+\lambda H_x(\cdot))} B_2(\gamma_x) \quad \text{in} \quad \ell^\infty[0,T]. \end{aligned}$$

Let

$$\begin{aligned} \eta_x(c) &= P\left(\sup_{0\le t\le T} \left| B_1(L_x(t)) + \frac{\lambda^{1/2}\varphi_x(\bar{H}_x(t))}{\gamma_x\varphi_x'(H_x(t))(H_x(t)+\lambda H_x(t))} B_2(\gamma_x) \right| \le c \right) \\ \eta_{xh}(c) &= P\left(\sup_{0\le t\le T} \left| B_1(L_{xh}(t)) + \frac{\lambda^{1/2}\varphi_x(\bar{H}_{xh}(t))}{\gamma_{xh}\varphi_x'(H_{xh}(t))(H_{xh}(t)+\lambda H_{xh}(t))} B_2(\gamma_{xh}) \right| \le c \right). \end{aligned}$$

Since $\sup_{0\le t\le T}|\cdot|$ is a continuous functional, we have that as $n \to +\infty$, $\eta_{xh}(c) \to \eta_x(c)$ for all c. By Lemma 4.4 below, we have that $\eta_x(\cdot)$ is a continuous function, and hence $\sup_{c>0}|\eta_{xh}(c) - \eta_x(c)| \to 0$ by Pólya's Theorem (see for example, the work of Serfling (1980)). More specifically, we see that $\eta_{xh}(c_{x\alpha h}) - \eta_x(c_{x\alpha h}) \to 0$ and by the definition of $c_{x\alpha h}$ we get that $\eta_x(c_{x\alpha h}) \to 1-\alpha$ which finishes our proof. □

Lemma 4.4. *Let $\{B_1(s)|0 \le s \le 1\}$ and $\{B_2(s)|0 \le s \le 1\}$ be independent Brownian bridges. Let $L_x(t),\ (0 \le t \le T)$ be as in (4.5), $\lambda > 0$. Then*

$$\sup_{0\le t\le T} \left| B_1(L_x(t)) + \frac{\lambda^{1/2}\varphi_x(\bar{H}_x(t))}{\gamma_x\varphi_x'(H_x(t))(H_x(t)+\lambda H_x(t))} B_2(\gamma_x) \right|$$

has a continuous distribution.

We omit the proof of this lemma since it follows the exact lines as the proof of Lemma A.4 of Van Keilegom and Veraverbeke (1997b), if we take

$$Y_x(t) = B_1(L_x(t)) + \frac{\lambda^{1/2}\varphi_x(\bar{H}_x(t))}{\gamma_x\varphi_x'(H_x(t))(H_x(t)+\lambda H_x(t))}B_2(\gamma_x)$$

4.4 A simulation study

In this section we perform a simulation study to investigate the finite sample coverage probability of the asymptotic confidence band of Theorem 4.2. The covariance structure of the limiting process $W(.|x)$ in Theorem 4.1 precludes the possibility to readily find values of $c_{\alpha xh}$ to satisfy (4.4). As a consequence, exact confidence bands for $F_x(t)$ cannot be obtained. To circumvent this problem, we develop in this section, an asymptotically conservative confidence band. Therefore we start with the fact that the left-hand side of (4.4) satisfies the inequality

$$P\left(\sup_{0\le t\le T}|B_1(L_{xh}(t))| + \sup_{0\le t\le T}\left|\frac{\lambda^{1/2}\varphi_x(\bar{H}_{xh}(t))}{\gamma_{xh}\varphi_x'(H_{xh}(t))(H_{xh}(t)+\lambda H_{xh}(t))}B_2(\gamma_{xh})\right| \le c_{\alpha xh}\right)$$
$$\le P\left(\sup_{0\le t\le T}\left|B_1(L_{xh}(t)) + \frac{\lambda^{1/2}\varphi_x(\bar{H}_{xh}(t))}{\gamma_{xh}\varphi_x'(H_{xh}(t))(H_{xh}(t)+\lambda H_{xh}(t))}B_2(\gamma_{xh})\right| \le c_{\alpha xh}\right) \quad (4.6)$$

Using the independence of $B_1(L_{xh}(t))$ and $B_2(\gamma_{xh})$, we convolve and rewrite the left-hand side of (4.6) as

$$\int_0^{c_{\alpha xh}} P\left(\sup_{0\le t\le T}|B_1(L_{xh}(t))| \le c_{\alpha xh}-y\right) dP\left(||A|| \le y\right)$$
$$= \int_0^{c_{\alpha xh}} Q_{d_{xh}(T)}\left(c_{\alpha xh}-y\right) dP\left(|N| \le \frac{y}{M_{xh}(\gamma_{xh},L_{xh}(T),\lambda)}\right) \quad (4.7)$$

where
$||A|| = \sup_{0\le t\le T}\left|\frac{\lambda^{1/2}\varphi(\bar{H}_{xh}(t))}{\gamma_{xh}\varphi\prime(H_{xh}(t))(H_{xh}(t)+\lambda H_{xh}(t))}B_2(\gamma_{xh})\right|$, $d_{xh}(T) = \frac{L_{xh}(T)}{1-L_{xh}(T)}$,
$M_{xh}(\gamma_{xh},L_{xh}(T),\lambda) = (\lambda\beta_{xh})^{(1/2)}\sup_{0\le t\le T}\left|\frac{\varphi(\bar{H}_{xh}(t))(1-L_{xh}(t,\lambda))}{\varphi_x\prime(H_{xh}(t))H_{xh}(t)}\right|$, $\beta_{xh} = \frac{1-\gamma_{xh}}{\gamma_{xh}}$ and N denotes a standard normal random variable.

Mimicking Hollander and Pẽna (1989), we define a distribution function

$$Q^{\star}(c_{\alpha x},\gamma_x,L_x(T),\lambda) = \sqrt{\frac{2}{\pi}}\frac{1}{M_x(c_{\alpha x},\gamma_x,L_x(T),\lambda)}\times \int_0^{c_{\alpha x}} Q_{d_x(T)}(c_{\alpha x}-y)\exp\left(-\frac{1}{2}\left(\frac{y}{M_x(\gamma_x,L_x(T),\lambda)}\right)^2\right)dy$$

where $d_x(T)=\frac{L_x(T)}{1-L_x(T)}$ and $Q_{d_x(T)}$ is defined as

$$Q_{d_x(T)}(c_x) = 1-2\Phi\left(-c_x\frac{1+d_x(T)}{d_x(T)^{1/2}}\right)+2\sum_{k=1}^{\infty}(-1)^k\exp\left(-2c_x^2k^2\right)\times \left\{\Phi\left(c_x\frac{d_x(T)+2k+1}{d_x(T)^{1/2}}\right)-\Phi\left(-c_x\frac{d_x(T)-2k+1}{d_x(T)^{1/2}}\right)\right\}$$

with $\Phi(.)$ being the standard normal cumulative distribution function. By choosing $c_{\alpha x}$ to satisfy $Q^{\star}(c_{\alpha x},\gamma_x,L_x(T),\lambda)=1-\alpha$, we obtain an asymptotically conservative confidence band

$$P\left[F_{xh}(t)-c_{\alpha xh}D_{xh}(t)\leq F_x(t)\leq F_{xh}(t)+c_{\alpha xh}D_{xh}(t)\right]\geq 1-\alpha \tag{4.8}$$

To investigate the coverage probabilities of (4.8), we generate data by taking fixed and equidistant design points $x_i=\frac{i}{n}$ $(i=1,2,3,\ldots,n)$. Also, we assume that the survival times Y_i $(i=1,2,3,...,n)$ are independent random variables with $Y_i\sim$ Weibull$(a_1+a_2x_i,b)$ such that for each design point the conditional survival function $\bar{F_i}(t)$ is given as

$$\bar{F_i}(t)=\exp\left(-\left(\frac{t}{b}\right)^{(a_1+a_2x_i)}\right)$$

for some constants a_1,a_2 such that $a_1>\wedge(0,-a_2)$ and $b>0$. Note that $a_1+a_2x_i$ characterizes the shape of the survival distribution of the ith subject whereas b is the scale parameter.

Furthermore, we assume that the censoring intensity parameter $\beta_{x_i}=\exp(a_3+a_4x_i)$ $(i=1,2,3,\ldots n)$ for some constants a_3 and a_4. Using the relation

$$\bar{G_i}(t) = \varphi_x^{[-1]}\left(\beta_{x_i}\varphi_x(\bar{F_i}(t))\right),$$

we obtain informative censoring times C_i based on the Clayton and Frank copula generator functions $\varphi_x(\cdot)$ at a pre-specified covariate level x with dependence parameter θ as follows:

1. we generate two independent uniform (0,1) random variables u and t.
2. we set $v = c_u^{-1}(t)$, where $c_u(v) = \frac{\partial}{\partial u}\left\{\varphi_x^{(-1)}\left(\varphi_x(u)+\varphi_x(v)\right)\right\}$ and c_u^{-1} is the inverse or quasi-inverse of c_u depending on whether φ_x is a strict or non-strict generator function.
3. we set $C_i = \overline{G}_i^{(-1)}(v)$ and $Y_i = \overline{F}_i^{(-1)}(u)$.

In particular, we use generators $\varphi_x(t) = \frac{1}{\theta}\left(t^{-\theta}-1\right)$ and $\varphi_x(t) = -\log\left(\frac{\exp(-\theta t)-1}{\exp(-\theta)-1}\right)$ for the Clayton and Frank copulas respectively. We investigate the effect of the association structure on the coverage probabilities by considering different choices of θ. Note that each choice of θ will lead to a different dependence structure for the Clayton and Frank copulas. Therefore, we use Kendall's τ as a measure of dependence so as to compare results under the two copula families. This dependence measure is defined as

$$\tau(x) \quad = \quad 1+4\int_0^1 \frac{\varphi_x(t)}{\varphi_x'(t)}dt$$

in Nelsen (2006) such that $-1 \leq \tau(x) \leq 1$, where the dependence gets stronger as $\tau(x)$ goes away from zero. Also, we investigate the effect of the censoring intensity on the coverage probabilities. That is, for each value of $\tau(x)$, we study three different sets of parameters a_1, a_2, a_3 and a_4. In the first set $(a_1 = 1, a_2 = 0.5, a_3 = -2.2, a_4 = 2)$, we chose the parameters such that the percentage of censored observations is always smaller than 45% (i.e. light censoring). In the second set $(a_1 = 1, a_2 = 0.5, a_3 = -0.2, a_4 = 0.4)$, the percentage of censored observations is inclusively between 45 and 55% (i.e. medium censoring); whereas in the third set $(a_1 = 1, a_2 = 0.5, a_3 = 0.2, a_4 = 0.5)$, the parameters are such that the percentage of censored observations is always greater than 55% (i.e. heavy censoring). At each combination of parameters, we generate 2000 samples, each of a size n. For each of these samples, we estimate the conditional Koziol-Green survival distribution at a pre-specified covariate level x together with the corresponding 95% confidence band. We use the Gasser-Müller weights given in Section 1.2 with the biquadratic kernel $K(z) = (15/16)\left(1-z^2\right) I\left(|z| \leq 1\right)$, since it is the most used type of weights in fixed design settings. Also, we use bandwidth $h_n = \left(\log n/n^{3/2}\right)^{2/11}$ so that as $n \to +\infty$, $nh_n \to 0$ and $(nh_n)^{-1/2}\log n \to 0$. Note that this bandwidth is based on the assumption made in Theorem 4.2.

Table 4.1: Coverage probabilities of the asymptotic confidence band at covariate levels of 0.65 and 0.97 using the Clayton copula

		Coverage (%)					
		Clayton ($\lambda = 1$)			Clayton ($\lambda = \gamma_x^2$)		
Dependence	Nominal (%)	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3
Covariate level = 0.65							
	90.0	97.9	99.5	99.5	99.8	99.9	99.9
$\tau = -0.99$	95.0	99.0	99.9	99.9	99.9	99.9	99.9
	99.0	99.9	99.9	99.9	99.9	99.9	99.9
	90.0	98.1	98.7	99.5	99.8	99.9	99.9
$\tau = 0.00$	95.0	99.4	99.1	99.9	99.9	99.9	99.9
	99.0	99.9	99.9	99.9	99.9	99.9	99.9
	90.0	94.9	94.3	94.2	99.7	99.9	99.9
$\tau = 0.99$	95.0	98.2	97.5	96.4	99.9	99.9	99.9
	99.0	99.6	99.6	99.2	99.9	99.9	99.9
Covariate level = 0.65							
	90.0	87.9	94.4	94.7	98.7	99.7	99.8
$\tau = -0.99$	95.0	91.8	98.5	96.8	99.5	99.9	99.9
	99.0	97.9	99.5	99.3	99.8	99.9	99.9
	90.0	87.5	93.5	93.5	98.4	99.5	99.9
$\tau = 0.00$	95.0	93.2	96.5	97.6	99.7	99.8	99.9
	99.0	96.4	99.1	99.5	99.8	99.9	99.9
	90.0	82.6	78.5	74.6	98.4	99.7	99.8
$\tau = 0.99$	95.0	86.3	85.6	82.1	98.8	99.8	99.9
	99.0	94.8	94.8	92.5	99.8	99.9	99.9

Next, we compute the coverage probability as the percentage of samples for which the confidence band at x covers its corresponding true survival distribution. In particular, we consider estimation at $x = 0.97$ and $x = 0.65$ as extreme and non-extreme covariate levels respectively in order to get some insight into the effect of x on the coverage probabilities. Also, we consider the cases $\lambda = 1$ and $\lambda = \gamma_x^2$ so as to obtain less and more conservative confidence bands respectively. In addition, we repeat the above process for different values of n (i.e. $n = 20, 30, 50, 100, 200, 300$) so as to examine also, the influence of n

on the coverage probabilities. Nevertheless, we report only results corresponding to the minimum sample size (i.e. $n = 50$) for which the coverage probabilities (at extreme or non-extreme covariate level) are at least their corresponding nominal confidence level. Note that the results for $\tau = 0$ are only given in Table 4.1 since it represents the independent copula which is a special case for both the Clayton and Frank copula when $\theta \to 0$.

Table 4.2: Coverage probabilities of the asymptotic confidence band at covariate levels of 0.65 and 0.97 using the Frank copula

		Coverage (%)					
		Frank ($\lambda = 1$)			Frank ($\lambda = \gamma_x^2$)		
Dependence	Nominal (%)	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3
Covariate level = 0.65							
	90.0	99.1	99.6	99.7	99.9	99.9	99.9
$\tau = -0.99$	95.0	99.4	99.8	99.8	99.9	99.9	99.9
	99.0	99.9	99.9	99.9	99.9	99.9	99.9
	90.0	96.1	94.8	92.9	99.6	99.9	99.9
$\tau = 0.99$	95.0	97.8	98.1	96.5	99.9	99.9	99.9
	99.0	99.8	99.5	99.4	99.9	99.9	99.9
Covariate level = 0.65							
	90.0	89.6	96.3	94.1	98.6	99.6	99.9
$\tau = -0.99$	95.0	93.0	98.1	98.5	99.6	99.9	99.9
	99.0	97.1	99.4	99.4	99.8	99.9	99.9
	90.0	80.6	77.2	75.2	98.3	99.4	99.9
$\tau = 0.99$	95.0	88.1	84.9	82.8	98.9	99.7	99.9
	99.0	95.8	94.5	92.7	99.9	99.9	99.9

In Tables 4.1 and 4.2 we observe that use of Clayton and Frank copulas results in similar coverage probabilities at equivalent censoring intensities and dependence structures. This implies that the choice of the copula function (i.e. Clayton or Frank) does not have a significant influence on the coverage probabilities. However, assuming $\lambda = \gamma_x^2$, leads to a non-decreasing trend in the coverage probability with increasing censoring intensity. This can be explained (at least in part) by the fact that as censoring increases, the

rate of deviation of the conditional Koziol-Green survival function estimate from the true survival function is negligible compared to the rate at which the bands increase with increasing censoring.

Furthermore, we observe at the extreme covariate level that, the coverage probabilities are at least their corresponding nominal only when we assume $\lambda = \gamma_x^2$. In contrast, the coverage probabilities at the non-extreme covariate level are always at least their corresponding nominal irrespective of whether we assume $\lambda = \gamma_x^2$ or $\lambda = 1$. Also for the non-extreme covariate level, assuming $\lambda = 1$ results in coverage probabilities which are at most those under the assumption that $\lambda = \gamma_x^2$. As already mentioned, assuming $\lambda = 1$ yields less (relative to $\lambda = \gamma_x^2$) conservative confidence bands. As such, the particular choice of λ depends on whether one wants a less conservative confidence band.

4.5 Real data illustration: Worcester heart attack study

In this section, we illustrate the asymptotic conditional Koziol-Green confidence band on a real data set. The data set comes from the Worcester Heart Attack Study (WHAS) which was introduced in Chapter 1. As mentioned there, this data set has information on more than 8000 admissions. Nonetheless, we only consider the 10% random sample of the original data set presented by Hosmer and Lemeshow (1999). As a consequence, the data set we utilize in the section has information on only 481 patients. Of these patients, 82 (17%) died while in admission (censored) whereas 399 (83%) were discharged (uncensored). We will mainly be concerned about the time until discharge from hospital of such patients. It is worth pointing out that, the results of this section are only for illustrative purpose. As such, we do not give a comparison with respect to the analysis of the complete data set. For details and pointers towards the findings from the complete WHAS data set, we refer to Hosmer and Lemeshow (1999).

In this study, we observe that a patient with severe health condition is likely to die within the first few days of admission. However, if such patient does not die, then he/she is most likely to spend many days in hospital bed. Not only severe health conditions would increase the days that a patient spends in the hospital, but also, for example, an infection

from the hospital can increase his/her days in the hospital bed. As such, we allege that time until discharge from hospital Y_i of a patient depends on the time until death in the hospital C_i (i.e. time until discharge has a negative influence on the time until death in the hospital).

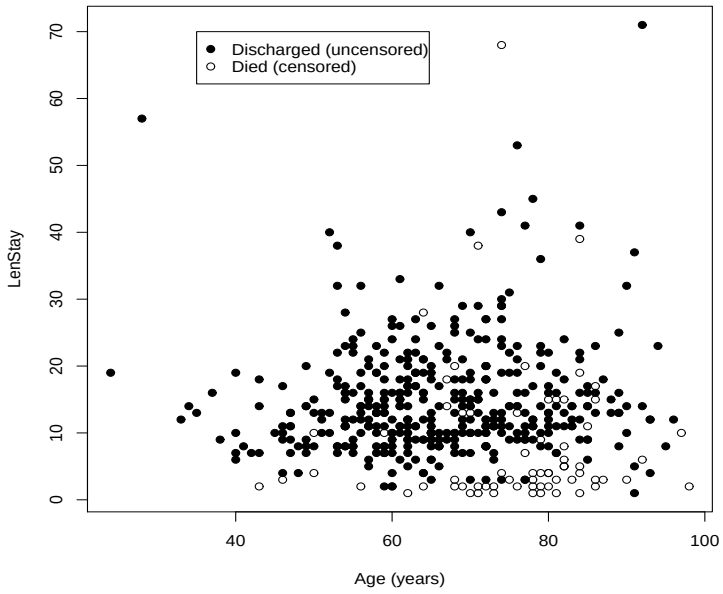

Figure 4.2: Scatter plot of time spent in hospital (LenStay) versus Age.

Figure 4.2 is a scatter plot of the observed time spent in hospital (LenStay) verses age of the patient at admission (Age) with a distinction between censored and uncensored patients. From the figure, we observe that most of the censored observations occurred among patients whose age is in the neighborhood of 80 years. This suggests possible association between censoring time and age of patients at admission. To formally investigate the applicability of the conditional Koziol-Green model, we adapt the partial Koziol-Green goodness-of-fit test of Braekers and Veraverbeke (2003) and calculate the Kolmogorov-Smirnov, the Cramer-von Mises and the Anderson-Darling types of test

statistics given respectively as

$$\begin{aligned} K_{nx} &= \left(\frac{nh_n}{||K||_2^2\gamma_{xh}(1-\gamma_{xh})}\right)^{1/2} \max_{1\le i\le n-1}|V_{n,i}^1 - V_{n,i}| \\ W_{nx}^2 &= \frac{nh_n}{||K||_2^2\gamma_{xh}(1-\gamma_{xh})}\sum_{i=1}^{n-1}(V_{n,i}^1 - \gamma_{xh}V_{n,i})^2 w_{n(i)}(x;h_n) \\ A_{nx}^2 &= \frac{nh_n}{||K||_2^2\gamma_{xh}(1-\gamma_{xh})}\sum_{i=1}^{n-1}\frac{(V_{n,i}^1 - \gamma_{xh}V_{n,i})^2}{V_{n,i}(1-V_{n,i})} w_{n(i)}(x;h_n) \end{aligned}$$

with $||K||_2^2 = \frac{5}{7}$, $V_{n,i}^1 = \sum_{k=1}^{i} w_{n(k)}(x;h_n)I(\delta_{(k)}=1)$ and $V_{n,i} = \sum_{k=1}^{i} w_{n(k)}(x;h_n)$, $(i = 1,2,...n = 481)$ where $\delta(k)$ and $w_{n(k)}(x;h_n)$ denotes respectively, the censoring indicator and Gasser-Müller weights (with the biquadratic kernel) corresponding to the ordered observed time spent in the hospital. We test at ages 50 and 75 years (i.e. $x = 50$ and 75). Hereby we take as bandwidth, $h_n = 43$. This choice is only to illustrate our method. We considered other choices $h_n = 33$ and $h_n = 53$ (not shown) but they gave similar results. A formal method to find the optimal bandwidth is a research area which we do not pursue in this thesis, but it could be a topic of future research.

Table 4.3: Conditional Koziol-Green goodness-of-fit test at ages 50 and 75 years

Age (years)	50		75	
	Statistic	P-Value	Statistic	P-Value
Kolmogorov-Smirnov	0.5536	0.9191	1.0033	0.2664
Cramer-von Mises	0.0735	0.7213	0.2934	0.1396
Anderson-Darling	0.8386	0.4531	2.4294	0.0689

From Table 4.3, we observe that the *p*-values associated with the three goodness-of-fit test statistics are larger than 5% (critical level). Thus, we fail to reject the conditional independence of the Z_x and the δ_x. Therefore, we allege that the conditional Koziol-Green model may be appropriate for the data set at 50 and 75 years. Using the Clayton and Frank copulas on this data set, we construct and compare confidence bands around the conditional Koziol-Green estimate of the survival (length of stay in hospital) function at ages 50 (middle aged patients) and 75 years (elderly patients). In the sequel, we assume $\lambda = 1$ so as to obtain less conservative (relative to $\lambda = \gamma_x^2$) confidence bands. In

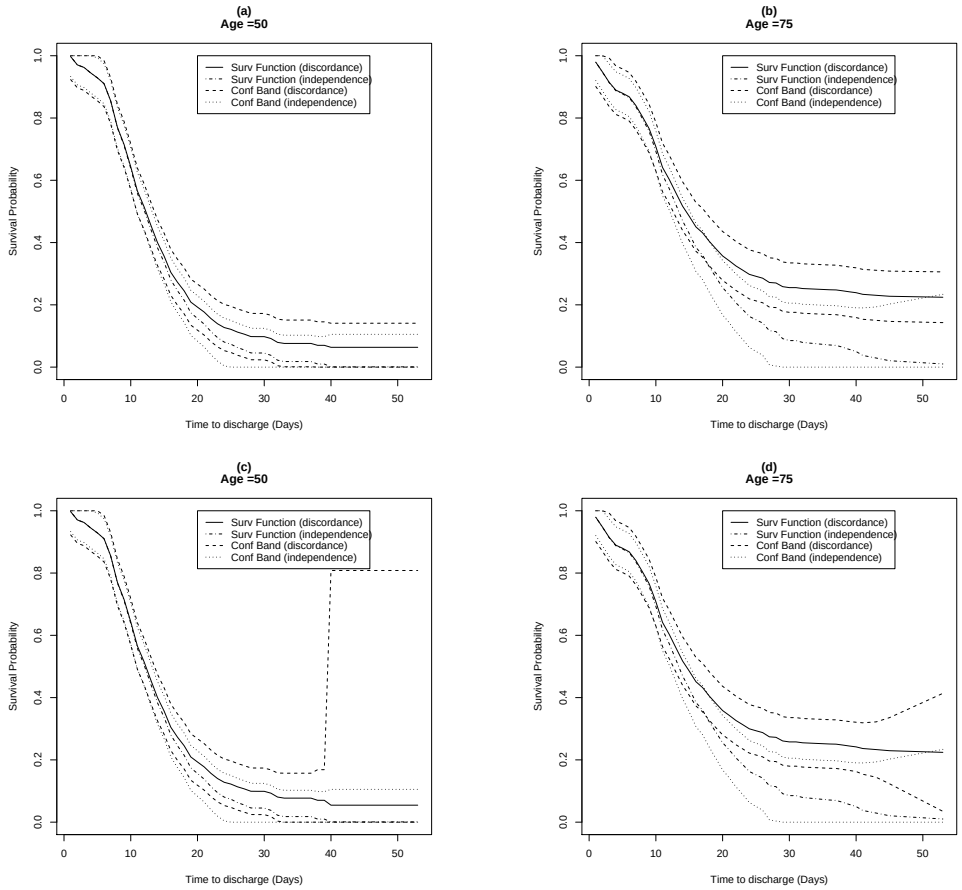

Figure 4.3: The conditional Koziol-Green survival function estimates (Surv Function) and associated 95% confidence bands (Conf Band) for middle aged (age = 50 years) and elderly (age = 75 years) patients under the Clayton (a & b) and Frank (c & d) copulas.

addition, we again use the Gasser-Müller weights with the biquadratic kernel and bandwidth $h_n = 43$. Figure 4.3 is a graphical representation of the conditional Koziol-Green survival distribution at ages 50 and 75 years for the AMI patients together with their corresponding 95% confidence band. In the figure, we consider two different association structures between the survival time (i.e time until discharge) and the censoring time (i.e. time until death in the hospital). Firstly, we assume that the survival time and censoring time are discordant (i.e. $\tau = -0.99$) since we expect that small death times in the hospital are related to large discharge times and vice versa. For a formal definition

of discordance, we refer to the book by Nelsen (2006). Secondly, we assume that the discharge time and time until death in the hospital are independent (i.e. $\tau = 0$). Note that the later assumption may be wrong for this data set. However, it is commonly used in other real data analyses. Therefore, we consider this choice only as reference for comparison with the result under the discordant association.

At 50 years, we observe under the Clayton and Frank copulas (Figure 4.3) that the survival distribution under the independent and discordant associations are close to each other. As a result, the confidence band constructed under the independent association clearly covers the survival distribution under the discordant association, and vice versa. This means that, ignoring the possibility of a dependence between the time until discharge from the hospital and the time until death in the hospital may not have any significant influence on the estimates based on the conditional Koziol-Green survival function and its associated 95% confidence band for middle aged patients. However, the same cannot be said about elderly patients since Figure 4.3 (i.e. (b) and (d)) indicate that the estimated survival distributions under independent and discordant associations at 75 years are clearly separated from each other; and that the confidence band under one form of association does not consistently cover the survival function under the other form of association.

5

The generalized conditional Koziol-Green model under dependent censoring

In Chapter 1, we introduced the conditional Koziol-Green estimator, which was pioneered by Braekers and Veraverbeke (2008). Further, we studied this estimator and showed the weak convergence of the associated process in Chapter 4. As applications of the weak convergence result, we first showed the efficiency of the conditional Koziol-Green estimator over the copula graphics estimator of Braekers and Veraverbeke (2005). Secondly, we developed a confidence band for the estimator and obtained some numerical results.

An important feature of the conditional Koziol-Green estimator is that it assumes characterization (1.10), which holds if and only if the observable variables Z_x and δ_x are independent. In some situations however, it becomes necessary to allow for possible dependence between these variables. In view of this, we introduced a generalization of the conditional Koziol-Green estimator in Section 1.2.2 of Chapter 1. In the present chapter, we study further this generalized estimator and obtain some associated attractive theoretical and numerical results. To be precise, we establish the strong consistency of the generalized conditional Koziol-Green estimator in Section 5.1. In Section 5.2, we give an asymptotic almost sure representation of the estimator which, as in the previous chapters paves way for establishing the weak convergence of the associated process in Section 5.3. Further, we investigate the finite sample performance of the estimator via a simulation study in Section 5.4. We conclude the chapter with an illustration of the estimator on the survival of Atlantic halibut data set.

Before giving these results, we complement the definitions and regularity assumptions given in Chapter 4 with the following:

Notations:

1. $\dot{\gamma}_x = \frac{d}{dx}\gamma_x$, $\ddot{\gamma}_x = \frac{d^2}{dx^2}\gamma_x$, $\dot{H}_x(t) = \frac{\partial}{\partial x}H_x(t)$, $\ddot{H}_x(t) = \frac{\partial^2}{\partial x^2}H_x(t)$

2. $||\dot{\gamma}_x|| = \sup_{x\in[0,1]}|\dot{\gamma}_x|$, $||\ddot{\gamma}_x|| = \sup_{x\in[0,1]}|\ddot{\gamma}_x|$, $||\dot{H}_x|| = \sup_{x\in[0,1]}\sup_{t\in[0,T]}|\dot{H}_x(t)|$, $||\ddot{H}_x|| = \sup_{x\in[0,1]}\sup_{t\in[0,T]}|\ddot{H}_x(t)|$

3. For some general copula function $\mathscr{C}_x(\cdot,\cdot)$, we let $\mathscr{C}_{x,ij}(u,v) = \frac{\partial^{i+j}}{\partial u^i \partial v^j}\mathscr{C}(u,v)$ denote the *i*th and *j*th partial derivatives with respect to its first and second coordinates respectively

Assumption:

(C6) At every design point $x \in [0,1]$ and for every $u \in (0,1)$, the derivatives $\mathscr{C}_{x,02}(u,v)$, $\mathscr{C}_{x,20}(u,v)$ and $\mathscr{C}_{x,11}(u,v)$ exist and are continuous for all $v \in [0,1]$.

We do not provide a discussion of this assumption, because it is of the same nature as Assumption (*A*1) in Chapter 2 and can also be verified accordingly.

5.1 Strong consistency result

The main result of this section is the strong consistency of the generalized conditional Koziol-Green estimator of the survival distribution function, as given in (1.25). Nevertheless, we also obtain an exponential inequality for the estimator. We formalize these as Theorem 4.1 whose proof relies on Lemma 5.1 below. We omit the proof of this lemma, since it follows the same lines as that of Lemma 2.1.

Lemma 5.1. *If* $\zeta \geq 0$, $0 \leq \eta < 1 - H_x(T)$ *and* $\eta = \frac{\varphi_x'(1)\zeta}{2\varphi_x'(1-H_x(T)-\eta)}$, *then* $\forall T \leq T_{H_x}$

$$\frac{\varphi_x'(1)\zeta}{2\varphi_x'\left(\varphi_x^{-1}\left(\varphi_x(1-H_x(T)) - \varphi_x'(1)\frac{\zeta}{2}\right)\right)} \leq \eta \leq 1 - H_x(T) - \varphi_x^{-1}\left(\varphi_x(1-H_x(T)) - \varphi_x'(1)\frac{\zeta}{2}\right),$$

for ζ *sufficiently small.*

Theorem 5.1. *Under Conditions (C1)-(C6), suppose* $T < T_{H_x}$, $\varphi_x'(1) < 0$ *and* $h_n \to 0$ *as* $n \to \infty$.

(a) *For* $\varepsilon > 0$ *and n sufficiently large, we have*

$$P\left(\sup_{0\leq t\leq T} |F_{xh}(t) - F_x(t)| > \varepsilon\right) \leq 2\exp(-d_1 n h_n \alpha_x^2 \varepsilon^2) + d_2 n h_n \beta_x \exp\left(\frac{-d_3 n h_n \alpha_x^2 \varepsilon^2}{4}\right)$$

with

$$\begin{aligned} \alpha_x &= \frac{\varphi_x'(1)}{2\varphi_x'\left(\varphi^{-1}\left(\varphi_x(1-H_x(T)) - \varphi_x'(1)\frac{\varepsilon}{2}\right)\right)}, \\ \beta_x &= 1 - H_x(T) - \varphi^{-1}\left(\varphi_x(1-H_x(T)) - \varphi_x'(1)\frac{\varepsilon}{2}\right) \end{aligned}$$

and d_1, d_2, d_3 *denoting finite positive constants.*

(b) *If* $(nh_n)^{-1}\log n \to 0$, *then*

$$\sup_{0\leq t\leq T} |F_{xh}(t) - F_x(t)| \to 0 \quad a.s.$$

Proof. The proof of this theorem is similar to that of Theorem 2.1. But for completeness, we repeat the lines here. Thus, we use the mean value theorem to obtain

$$\begin{aligned} F_{xh}(t)-F_x(t) &= -\left[\varphi_x^{-1}\left(-\int_0^{H_{xh}(t)}\varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_{xh},w)dw\right)\right. \\ &\quad \left.-\varphi_x^{-1}\left(-\int_0^{H_x(t)}\varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_x,w)dw\right)\right] \\ &= A(\gamma^*,H^*(t))(\gamma_{xh}-\gamma_x)+B(\gamma^*,H^*(t))(H_{xh}(t)-H_x(t)) \end{aligned}$$

where

$$A(\gamma^*,H^*(t)) = \frac{\int_0^{H^*(t)}\varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma^*,w)dw}{\varphi_x'\left(\varphi_x^{-1}\left(-\int_0^{H^*(t)}\varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma^*,w)dw\right)\right)}$$

and

$$B(\gamma^*,H^*(t)) = \frac{\varphi_x'(1-H^*(t))\mathscr{C}_{x,01}(\gamma^*,H^*(t))}{\varphi_x'\left(\varphi_x^{-1}\left(-\int_0^{H_x(t)}\varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_x,w)dw\right)\right)}$$

with γ^* between γ_{xh} and γ_x, $H^*(t)$ between $H_{xh}(t)$ and $H_x(t)$.Using integration by parts, we can easily show that

$$\sup_{0\le t\le T}|A(\gamma^*,H^*(t))| \le \frac{3}{|\varphi_x'(1)|}\sup_{0\le t\le T}|\varphi_x'(1-H^*(t))|$$

and

$$\sup_{0\le t\le T}|B(\gamma^*,H^*(t))| \le \frac{1}{|\varphi_x'(1)|}\sup_{0\le t\le T}|\varphi_x'(1-H^*(t))|$$

Therefore, for all $\varepsilon>0$

$$\begin{aligned} &P\left(\sup_{0\le t\le T}|F_{xh}(t)-F_x(t)|>\varepsilon\right) \\ &\quad\le P\left(|\gamma_{xh}-\gamma_x|\sup_{0\le t\le T}|A(\gamma^*,H^*(t))|>\frac{\varepsilon}{2}\right) \\ &\qquad +P\left(\sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|\sup_{0\le t\le T}|B(\gamma^*,H^*(t))|>\frac{\varepsilon}{2}\right) \\ &\quad\le P\left(\frac{3}{|\varphi_x'(1)|}|\gamma_{xh}-\gamma_x|\sup_{0\le t\le T}|\varphi_x'(1-H^*(t))|>\frac{\varepsilon}{2}\right) \\ &\qquad +P\left(\frac{1}{|\varphi_x'(1)|}\sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|\sup_{0\le t\le T}|\varphi_x'(1-H^*(t))|>\frac{\varepsilon}{2}\right) \end{aligned}$$

For $\eta > 0$, we can write

$$
\begin{aligned}
&P\left(\sup_{0\le t\le T}|F_{xh}(t)-F_x(t)|>\varepsilon\right)\\
&\le P\left(\frac{3}{|\varphi_x(1)|}|\gamma_{xh}-\gamma_x|\sup_{0\le t\le T}|\varphi_x'(1-H^*(t))|>\frac{\varepsilon}{2},\ \sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|\le\eta\right)\\
&+P\left(\frac{3}{|\varphi_x(1)|}|\gamma_{xh}-\gamma_x|\sup_{0\le t\le T}|\varphi_x'(1-H^*(t))|>\frac{\varepsilon}{2},\ \sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|>\eta\right)\\
&+P\left(\frac{1}{|\varphi_x(1)|}\sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|\sup_{0\le t\le T}|\varphi_x'(1-H^*(t))|>\frac{\varepsilon}{2},\ \sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|\le\eta\right)\\
&+P\left(\frac{1}{|\varphi_x(1)|}\sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|\sup_{0\le t\le T}|\varphi_x'(1-H^*(t))|>\frac{\varepsilon}{2},\ \sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|>\eta\right)
\end{aligned}
$$

With $0<\eta<1-H_x(T)$ and satisfying

$$\sup_{0\le t\le T}|\varphi_x'(1-H^*(t))|<|\varphi_x'(1-H_x(T)-\eta)|,$$

we note that this is further bounded above by

$$
\begin{aligned}
&P\left(\sup_{0\le t\le T}|F_{xh}(t)-F_x(t)|>\varepsilon\right)\\
&\le P\left(|\gamma_{xh}-\gamma_x|>\frac{\varphi_x'(1)\varepsilon}{6\varphi_x'(1-H_x(T)-\eta)},\ \sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|\le\eta\right)\\
&+P\left(|\gamma_{xh}-\gamma_x|>\frac{\varphi_x'(1)\varepsilon}{6\varphi_x'(1-H_x(T)-\eta)},\ \sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|>\eta\right)\\
&+P\left(\sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|>\frac{\varphi_x'(1)\varepsilon}{2\varphi_x'(1-H_x(T)-\eta)},\ \sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|\le\eta\right)\\
&+P\left(\sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|>\frac{\varphi_x'(1)\varepsilon}{2\varphi_x'(1-H_x(T)-\eta)},\ \sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|>\eta\right)\\
&\le\ P\left(|\gamma_{xh}-\gamma_x|>\frac{\varphi_x'(1)\varepsilon}{6\varphi_x'(1-H_x(T)-\eta)}\right)\\
&+P\left(|\gamma_{xh}-\gamma_x|>\frac{\varphi_x'(1)\varepsilon}{6\varphi_x'(1-H_x(T)-\eta)},\ \sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|>\eta\right)\\
&+P\left(\sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|>\frac{\varphi_x'(1)\varepsilon}{2\varphi_x'(1-H_x(T)-\eta)}\right)\\
&+P\left(\sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|>\frac{\varphi_x'(1)\varepsilon}{2\varphi_x'(1-H_x(T)-\eta)},\ \sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|>\eta\right)
\end{aligned}
$$

Choosing η such that $\eta = \frac{\varphi_x'(1)\varepsilon}{2\varphi_x'(1-H_x(T)-\eta)}$, we obtain

$$\begin{aligned} P\left(\sup_{0\le t\le T}|F_{xh}(t)-F_x(t)|>\varepsilon\right) & \\ &\le P\left(|\gamma_{xh}-\gamma_x|>\frac{\eta}{3}\right)+3P\left(\sup_{0\le t\le T}|H_{xh}(t)-H_x(t)|>\eta\right) \\ &\le \exp(-\frac{d_1nh_n\eta^2}{9})+d_2nh_n\eta\exp\left(\frac{-d_3nh_n\eta^2}{4}\right) \\ &\le \exp(-\frac{d_1nh_n\alpha_x^2}{9})+d_2nh_n\beta_x\exp\left(\frac{-d_3nh_n\alpha_x^2}{4}\right) \end{aligned}$$

where

$$\begin{aligned} \alpha_x &= \frac{\varphi_x'(1)}{2\varphi_x'\left(\varphi^{-1}\left(\varphi_x(1-H_x(T))-\varphi_x'(1)\frac{\varepsilon}{2}\right)\right)}\ , \\ \beta_x &= 1-H_x(T)-\varphi^{-1}\left(\varphi_x(1-H_x(T))-\varphi_x'(1)\frac{\varepsilon}{2}\right) \end{aligned}$$

and d_1, d_2, d_3 are finite positive constants. In the preceding display, the second inequality follows from Braekers and Veraverbeke (2008) and requires the condition

$$\eta \ \ge\ \max\left(||\dot{\gamma}_x||_\infty\bar{\Delta}_n+||\ddot{\gamma}_x||_\infty\mu_2^K h_n^2, \sqrt{6}||K||_2(nh_n)^{-1/2}, 2||\ddot{H}_x||_\infty\mu_2^K h_n^2\right)$$

while the third inequality follows from Lemma 5.1 above. Taking

$$\varepsilon=\varepsilon_n=d_0(nh_n)^{-1/2}(\log n)^{1/2}$$

with d_0 a finite positive constant, we get the strong consistency result. □

5.2 Almost sure asymptotic representation

Similar to Chapters 2 and 3, we now present the generalized conditional Koziol-Green estimator as a weighted sum of n independent random variables. This representation, as mentioned before, is a device that aids in obtaining further theoretical properties. We formulate such a representation as the succeeding theorem and employ it to obtain a further property in Section 5.3.

Theorem 5.2. *Under Assumptions (C1)-(C4) and (C6), suppose $T < T_{H_x}$ and $h_n \to 0$. Then as $n \to \infty$,*

$$F_{xh}(t) - F_x(t) = \sum_{i=1}^{n} w_{n_i}(x,h_n) m_{tx}(Z_i,\delta_i) + r_n(t)$$

where $r_n(t) = O\left((nh_n)^{-1}\log n\right)$ *a.s., and*

$$\begin{aligned} m_{tx}(Z_i,\delta_i) = \frac{1}{\varphi_x'(F_x(t))}\Big\{ & (\mathbb{1}\{\delta_i = 1\} - \gamma_x)\int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x,w)dw \\ & + (\mathbb{1}\{Z_i \le t\} - H_x(t))\varphi_x'(1-H_x(t))\mathscr{C}_{x,01}(\gamma_x,H_x(t))\Big\} \end{aligned}$$

Proof. By a second order Taylor's expansion, we have

$$\begin{aligned} F_{xh}(t) - F_x(t) = \frac{1}{\varphi_x'(\bar{F}_x(t))}\Big\{ & \int_0^{H_{xh}(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_{xh},w)dw \\ & - \int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_x,w)dw\Big\} + r_{n_1}(t) \quad (5.1) \end{aligned}$$

where

$$\begin{aligned} r_{n_1}(t) = & \frac{\varphi_x''(\varphi_x^{-1}(\eta(t)))}{2(\varphi_x'(\varphi_x^{-1}(\eta(t))))^3} \times \\ & \left\{\int_0^{H_{xh}(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_{xh},w)dw - \int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_x,w)dw\right\}^2 \end{aligned}$$

with $\eta(t)$ between $-\int_0^{H_{xh}(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_{xh},w)dw$ and $-\int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_x,w)dw$.

Let

$$I(t) = \int_0^{H_{xh}(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_{xh},w)dw - \int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_x,w)dw.$$

Then,

$$\sup_{0\le t\le T}|r_{n_1}(t)| \le \frac{\varphi_x''(\varphi_x^{-1}(\eta(T)))}{2|\varphi_x'(1)|^3}\sup_{0\le t\le T}|I(t)|^2$$

and $\eta(T)$ lies between $-\int_0^{H_{xh}(T)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_{xh},w)dw$ and $-\int_0^{H_x(T)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_x,w)dw$. By Van Keilegom and Veraverbeke (1997a, Lemma A.2), we know that $H_{xh}(T) \to H_x(T)$

a.s. Thus, we may suppose that $T < T_{H_{xh}}$, since $T < T_{H_x}$. Let $H_x^{min}(T) = \min(H_{xh}(T), H_x(T))$ and $H_x^{max}(T) = \max(H_{xh}(T), H_x(T))$. Then it follows that the preceding inequality is further bounded above by

$$\frac{\varphi_x''(\varphi_x^{-1}(\eta^{oo}(T)))}{2|\varphi_x'(1)|^3} \sup_{0\le t\le T} |I(t)|^2 \quad ,$$

with $\eta^{oo}(T)$ given by

$$\eta^{oo}(T) = -\int_0^{H_x^{max}(T)} \varphi_x'(1-w)\mathscr{C}_{x,01}(1,w)dw.$$

By the mean value theorem, we further have that

$$\begin{aligned}
\sup_{0\le t\le T} |I(t)| &\le |\gamma_{xh} - \gamma_x| \sup_{0\le t\le T} \left| \int_0^{H^*(t)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma^*, w)dw \right| \\
&\quad + |\varphi_x'(1-H_x^*(T))| \sup_{0\le t\le T} |H_{xh}(t) - H_x(t)| \\
&\le |\varphi_x'(1-H_x^{max}(T))| \left\{ 4|\gamma_{xh} - \gamma_x| + |\sup_{0\le t\le T} |H_{xh}(t) - H_x(t)| \right\}
\end{aligned}$$

where the second inequality in the preceding display follows by integrating by parts, the first term at the right hand side of the first inequality in the same display.

From Van Keilegom and Veraverbeke (1997a, Lemma A.4) and Braekers and Veraverbeke (2001), it can respectively be shown that

$$\begin{aligned}
\sup_{0\le t\le T} |H_{xh}(t) - H_x(t)| &= O\left((nh_n)^{-1/2}(\log n)^{1/2}\right) \quad \text{a.s.} \ , \\
|\gamma_{xh} - \gamma_x| &= O\left((nh_n)^{-1/2}(\log n)^{1/2}\right) \quad \text{a.s.}
\end{aligned}$$

Consequently, we obtain

$$\sup_{0\le t\le T} |r_{n_1}(t)| \le \frac{\varphi_x''(\varphi_x^{-1}(\eta^{oo}(T)))}{2|\varphi_x'(1)|^3} \sup_{0\le t\le T} |I(t)|^2 = O\left(\frac{\log n}{nh_n}\right) \quad \text{a.s.} \qquad (5.2)$$

Next, we use a second order Taylor expansion and note that

$$\begin{aligned}
&\int_0^{H_{xh}(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_{xh}, w)dw - \int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_x, w)dw \\
&= (\gamma_{xh} - \gamma_x) \int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x, w)dw \\
&\quad + (H_{xh}(t) - H_x(t))\varphi_x'(1-H_x(t))\mathscr{C}_x(\gamma_x, H_x(t)) \qquad (5.3) \\
&\quad + r_{n_2}(t) + r_{n_3}(t) + r_{n_4}(t)
\end{aligned}$$

where

$$\begin{aligned} r_{n_2}(t) &= \frac{1}{2}(\gamma_{xh}-\gamma_x)^2 \int_0^{H_x^*(t)} \varphi_x'(1-w)\mathscr{C}_{x,21}(\gamma_x^*,w)dw \\ r_{n_3}(t) &= \frac{1}{2}(H_{xh}(t)-H_x(t))^2 \left\{\varphi_x'(1-H_x^*(t))\mathscr{C}_{x,02}(\gamma_x^*,H_x^*(t)) - \varphi_x''(1-H_x^*(t))\mathscr{C}_{x,01}(\gamma_x^*,H_x^*(t))\right\} \\ r_{n_4}(t) &= (\gamma_{xh}-\gamma_x)(H_{xh}(t)-H_x(t))\varphi_x'(1-H_x^*(t))\mathscr{C}_{x,11}(\gamma_x^*,H_x^*(t)) \end{aligned}$$

with γ_x^* between γ_{xh} and γ_x; and $H_x^*(t)$ between $H_{xh}(t)$ and $H_x(t)$. We now determine the rate of convergence of $r_{n_2}(t)$, $r_{n_3}(t)$ and $r_{n_4}(t)$.

Integrating by parts, we first obtain

$$\begin{aligned} \sup_{0\le t\le T}|r_{n_2}(t)| &\le |\gamma_{xh}-\gamma_x|^2\left\{\varphi_x'(1-H_x^*(T))\sup_{0\le t\le T}|\mathscr{C}_{x,20}(\gamma_x^*,H_x^*(t))|\right. \\ &\quad \left. + \sup_{0\le t\le T}\left|\int_0^{H_x^*(t)}\varphi_x''(1-w)\mathscr{C}_{x,20}(\gamma_x^*,w)dw\right|\right\} \\ &\le 3|\varphi_x'(1-H_x^*(T))|\sup_{0\le v\le 1}|\mathscr{C}_{x,20}(u,v)||\gamma_{xh}-\gamma_x| \ , \ \forall u\in(0,1) \\ &\le 3|\varphi_x'(1-H_x^{max}(T))|\sup_{0\le v\le 1}|\mathscr{C}_{x,20}(u,v)||\gamma_{xh}-\gamma_x| \ , \ \forall u\in(0,1) \end{aligned}$$

Using Assumption $(C4)$, we subsequently obtain

$$\sup_{0\le t\le T}|r_{n_2}(t)| = O\left(\frac{\log n}{nh_n}\right) \tag{5.4}$$

Analogously, we can easily show that

$$\sup_{0\le t\le T}|r_{n_3}(t)| = O\left(\frac{\log n}{nh_n}\right) \quad \text{and} \quad \sup_{0\le t\le T}|r_{n_4}(t)| = O\left(\frac{\log n}{nh_n}\right) \tag{5.5}$$

Substituting (5.2)and (5.3) into (5.1), preceded by (5.4) and (5.5) into (5.3) concludes the proof. □

5.3 Weak convergence result

As mentioned above, the essence of the almost sure asymptotic representation is to facilitate the derivation of further properties of the estimator under consideration. The objective of the present section is therefore, to use Theorem 5.2 and prove the weak

convergence of the process $(nh_n)^{1/2}(F_{xh}(\cdot)-F_x(\cdot))$ associated with the generalized conditional Koziol-Green estimator (1.25) in an appropriate space of functions that will become clear in what follows.

Theorem 5.3. *Suppose the conditions of Theorem 5.2 are satisfied.*

(a) *If in addition* $nh_n^5 \to 0$ *and* $n^{-1}(\log n)^3 \to 0$, *then*

$$(nh_n)^{1/2}(F_{xh}(\cdot)-F_x(\cdot)) \quad \to \quad W(\cdot|x)$$

where $W(\cdot|x)$ *is a zero mean Gaussian process with covariance function*

$$\begin{aligned}
&\Gamma_x(s,t)\\
&= \gamma_x(1-\gamma_x)\int_0^{H_x(s)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x,w)dw \int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x,w)dw\\
&+(H_x(s\wedge t)-H_x(s)H_x(t))\,\varphi_x'(1-H_x(s))\varphi_x'(1-H_x(t))\mathscr{C}_{x,01}(\gamma_x,H_x(s))\mathscr{C}_{x,01}(\gamma_x,H_x(t))\\
&+(H_x^u(t)-\gamma_x H_x(t))\varphi_x'(1-H_x(t))\mathscr{C}_{x,01}(\gamma_x,H_x(t))\int_0^{H_x(s)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x,w)dw\\
&+(H_x^u(s)-\gamma_x H_x(s))\varphi_x'(1-H_x(s))\mathscr{C}_{x,01}(\gamma_x,H_x(s))\int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x,w)dw
\end{aligned}$$

for all $s,t\in[0,T]$.

(b) *If* $h_n = cn^{-1/5}$, *for a positive constant c, then*

$$(nh_n)^{1/2}(F_{xh}(\cdot)-F_x(\cdot)) \quad \to \quad \widetilde{W}(\cdot|x)$$

where $\widetilde{W}(\cdot|x)$ *is a Gaussian process with variance-covariance function* $\Gamma_x(s,t)$ *and mean function* $B_x(\cdot)$ *given by*

$$\begin{aligned}
B_x(t) \quad = \quad & \frac{\mu_2^K h_n^2}{2\varphi_x'(F_x(t))}\left\{\ddot{\gamma}_x\int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x,w)dw\right.\\
& \left.+\ddot{H}_x(t)\varphi_x'(1-H_x(t))\mathscr{C}_x(\gamma_x,H_x(t))\right\}c^{5/2}
\end{aligned}$$

Proof. To show the weak convergence of the process $(nh_n)^{1/2}(F_{xh}(t)-F_x(t))$, we work

as in Braekers and Veraverbeke (2001). First, we define

$$\begin{aligned} b_n(x,t) &= \sum_{i=1}^{n} w_{n_i}(x;t) E\, m_{t_x}(Z_i,\delta_i) \\ &= \frac{1}{\varphi_x'(F_x(t))}\Big\{(\gamma_{x_i}-\gamma_x)\int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x,w)dw \\ &\quad +(H_{x_i}(t)-H_x(t))\varphi_x'(1-H_x(t))\mathscr{C}_x(\gamma_x,H_x(t))\Big\} \\ &= \frac{\mu_2^K h_n^2}{2\varphi_x'(F_x(t))}\Big\{\ddot{\gamma}_x\int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x,w)dw \\ &\quad +\ddot{H}_x(t)\varphi_x'(1-H_x(t))\mathscr{C}_x(\gamma_x,H_x(t))\Big\} + o(h_n^2) + O\left(n^{-1}\right) \end{aligned}$$

where the last equality in the preceding display follows from Aerts et al. (1994). This implies the bias

$$(nh_n)^{1/2} b_n(x,t) = \begin{cases} o(1) & \text{if} \quad nh_n^5 \to 0 \\ B_x(t) & \text{if} \quad h_n = cn^{-1/5} \end{cases}$$

where

$$\begin{aligned} B_x(t) &= \frac{\mu_2^K h_n^2}{2\varphi_x'(F_x(t))}\Big\{\ddot{\gamma}_x\int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x,w)dw \\ &\quad +\ddot{H}_x(t)\varphi_x'(1-H_x(t))\mathscr{C}_x(\gamma_x,H_x(t))\Big\}c^{5/2} \end{aligned}$$

with c denoting a finite positive constant. Therefore we write

$$\begin{aligned} F_{xh}(t)-F_x(t) &= \sum_{i=1}^{n} w_{n_i}(x;h_n)\left[m_{t_x}(Z_i,\delta_i)-E\, m_{t_x}(Z_i,\delta_i)\right] + b_n(x,t) + r_n(t) \\ &= \sum_{i=1}^{n} w_{n_i}(x;h_n)\xi_{t_x}(Z_i,\delta_i) + b_n(x,t) + r_n(t) \end{aligned} \tag{5.6}$$

Using Billingsley (1968, Theorem 4.1), it follows that the weak convergence of the process $(nh_n)^{1/2}(F_{xh}(\cdot)-F_x(\cdot))$ to a Gaussian process in the space of uniformly bounded real valued functions $\ell^\infty[0,T]$ is equivalent to the weak convergence of the process

$$W_{xh}(\cdot) = (nh_n)^{1/2}\sum_{i=1}^{n} w_{n_i}(x;h_n)\xi_{\cdot x}(Z_i,\delta_i)$$

to a Gaussian limit in $\ell^\infty[0,T]$. To do this, we first show the convergence of the finite dimensional distributions of the process and later append it with tightness in $\ell^\infty[0,T]$. For the convergence of the finite dimensional distributions, we define

$$W_{n_{k_i}} \quad = \quad (nh_n)^{1/2} W_{n_k}(x;h_n)\xi_{t_{ix}}(Z_k,\delta_k)$$

and verify the conditions of Araujo and Giné (1980), as before (see Chapter 3).

For the first condition (see Section 3.4.2), we find after some calculations that

$$\begin{aligned}
&E\left(W_{n_{k_i}} W_{n_{k_j}}\right) \\
&\quad = nh_n w_{n_k}^2(x;h_n) E\left[\xi_{t_{ix}}(Z_k,\delta_k)\xi_{t_{jx}}(Z_k,\delta_k)\right] \\
&\quad = nh_n w_{n_k}^2(x;h_n) \times \\
&\qquad \Big\{ \gamma_x(1-\gamma_x) \int_0^{H_x(t_i)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x,w)dw \int_0^{H_x(t_j)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x,w)dw \\
&\qquad + (H_x(t_i\wedge t_j) - H_x(t_i)H_x(t_j))\, \varphi_x'(1-H_x(t_i))\varphi_x'(1-H_x(t_j))\mathscr{C}_{x,01}(\gamma_x,H_x(t_i))\mathscr{C}_{x,01}(\gamma_x,H_x(t_j)) \\
&\qquad + (H_x^u(t_j) - \gamma_x H_x(t_j))\varphi_x'(1-H_x(t_j))\mathscr{C}_{x,01}(\gamma_x,H_x(t_j)) \int_0^{H_x(t_i)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x,w)dw \\
&\qquad + (H_x^u(t_i) - \gamma_x H_x(t_i))\varphi_x'(1-H_x(t_i))\mathscr{C}_{x,01}(\gamma_x,H_x(t_i)) \int_0^{H_x(t_j)} \varphi_x'(1-w)\mathscr{C}_{x,11}(\gamma_x,w)dw \Big\} \\
&\quad = \Gamma_x(t_i,t_j) \times nh_n w_{n_k}^2(x;h_n)
\end{aligned}$$

Subsequently, it follows from Lemma 3.1 of Van Keilegom and Veraverbeke (1997a) that

$$\lim_{n\to\infty} \sum_{k=1}^n E\left(W_{n_{k_i}} W_{n_{k_j}}\right) \quad = \quad \Gamma_x(t_i t_j) \times nh_n \sum_{k=1}^n w_{n_k}(x;h_n)^2 \quad = \quad \Gamma_x(t_i,t_j) + o(1).$$

For the second condition, we find after some calculations that

$$\begin{aligned}
\sup_{0\le t\le T} |\xi_{t_x}(Z_k,\delta_k)| \quad &\le \quad |\mathbb{1}\{\delta_k = 1\} - \gamma_{x_k}| \sup_{0\le t\le T} \left| \int_0^{H_x(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_x,w)dw \right| \\
&\qquad + \sup_{0\le t\le T} \left| (\mathbb{1}\{Z_k \le t\} - H_{x_k}(t))\varphi_x'(1-H_x(t))\mathscr{C}_{x,01}(\gamma_x,H_x(t)) \right| \\
&\le \quad \left|\varphi_x'(1-H_x(T))\right| \left(3 + \sup_{0\le t\le T} \mathscr{C}_{x,01}(\gamma_x,H_x(t))\right) < \infty
\end{aligned}$$

This gives that

$$\max_{1\le k\le n} |W_{n_k}| \quad = \quad O\left((nh_n)^{-1/2}\right) \qquad \text{and} \qquad \sum_{k=1}^n |W_{n_k}|^2 \quad = \quad O(1)$$

Consequently, we get

$$\begin{aligned}\sum_{k=1}^{n}\int_{\{|W_{n_k}|>\varepsilon\}}|W_{n_k}|^2dP &\leq \int_{\max_{1\leq k\leq n}|W_{n_k}|>\varepsilon}\sum_{k=1}^{n}|W_{n_k}|^2dP \\ &\leq O(1)P\left(\max_{1\leq k\leq n}|W_{n_k}|>\varepsilon\right)=o(1) \quad \text{as} \quad n\to\infty\end{aligned}$$

Hence, it follows from Araujo and Giné (1980) that $(W_{xh}(t_1),W_{xh}(t_2),...,W_{xh}(t_q))$ converges in distribution to $N(0,\Gamma_x(t_i,t_j))$ for any $q=1,2,...$ and any $0\leq t_1\leq t_2\leq\cdots\leq t_q$.

To establish tightness, we need to show that the process $W_{xh}(\cdot)=\sum_{i=1}^{n}(Z_{n_i}(\cdot)-EZ_{n_i}(\cdot))$ with $Z_{n_i}(t)=(nh_n)^{1/2}w_{n_i}(x;h_n)m_{t_x}(Z_i,\delta_i)$ is asymptotically tight in $\ell^\infty[0,T]$. This is equivalent to verifying the conditions of the bracketing central limit theorem of van der Vaart and Wellner (2000, Theorem 2.11.9), as given in Chapter 3.

To do so, we define the index set $\mathscr{F}=[0,T]$ and endowed with the semimetric ρ defined by

$$\rho(t,t') = \max\left(\begin{array}{l}\sup_{x\in[0,1]}\left|\dfrac{1}{\varphi_x'(F_x(t))}-\dfrac{1}{\varphi_x'(F_x(t'))}\right|,\ \sup_{x\in[0,1]}\left|\displaystyle\int_{H_x(t')}^{H_x(t)}\varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_x,w)dw\right|, \\ \sup_{x\in[0,1]}\left|\varphi_x'(1-H_x(t))\mathscr{C}_{x,01}(\gamma_x,H_x(t))-\varphi_x'(1-H_x(t'))\mathscr{C}_{x,01}(\gamma_x,H_x(t'))\right|, \\ \sup_{x\in[0,1]}|H_x(t)-H_x(t')|\end{array}\right)$$

for $t,t'\in\mathscr{F}$. $N_{[]}(\varepsilon,\mathscr{F},L_2^n)$ is the bracketing number for every n in a partition $\mathscr{F}=\cup_j\mathscr{F}_{\varepsilon_j}$ of the index set into sets $\mathscr{F}_{\varepsilon_j}$ such that

$$\sum_{i=1}^{n}E\left[\sup_{t,t'\in\mathscr{F}_{\varepsilon_j}}|Z_{n_i}(t)-Z_{n_i}(t')|^2\right] \leq \varepsilon^2 \quad , \quad \forall j=1,2,... \tag{5.7}$$

Before we check the first condition of the bracketing central limit theorem, we point out that the function $m_{t_x}(Z_i,\delta_i)$ is uniformly bounded above by

$$\sup_{t\in\mathscr{F}}|m_{t_x}(Z_i,\delta_i)| \leq |\varphi_x'(1-H_x(T)|\left\{2\mathscr{C}_{x,10}(\gamma_x,H_x(T))+\sup_{t\in\mathscr{F}}\mathscr{C}_{x,01}(\gamma_x,H_x(t))\right\}<\infty.$$

Consequently, we have

$$\begin{aligned}
\sup_{t\in\mathscr{F}}|Z_{n_i}(t)| \leq & (nh_n)^{1/2}|\varphi_x'(1-H_x(T)|\left\{2\mathscr{C}_{x,10}(\gamma_x,H_x(T))+\sup_{t\in\mathscr{F}}\mathscr{C}_{x,01}(\gamma_x,H_x(t))\right\}w_{n_i}(x:h_n)\\
\leq & (nh_n)^{1/2}|\varphi_x'(1-H_x(T)|\left\{2\mathscr{C}_{x,10}(\gamma_x,H_x(T))+\sup_{t\in\mathscr{F}}\mathscr{C}_{x,01}(\gamma_x,H_x(t))\right\}||K||_\infty O\left((nh_n)^{-1}\right)\\
\leq & |\varphi_x'(1-H_x(T)|\left\{2\mathscr{C}_{x,10}(\gamma_x,H_x(T))+\sup_{t\in\mathscr{F}}\mathscr{C}_{x,01}(\gamma_x,H_x(t))\right\}||K||_\infty O\left((nh_n)^{-1/2}\right)\\
= & O\left((nh_n)^{-1/2}\right)<\lambda
\end{aligned}$$

for sufficiently large n and for all $\lambda>0$. This implies, Condition 1 of the bracketing central limit theorem.

Because our partition $\mathscr{F}_{\varepsilon_j}$ of the index set is constructed independent of n, we do not need to verify the second condition since it is automatically satisfied. For the third condition, we divide the index set $\mathscr{F}=[0,T]$ into subintervals $[t_{j-1},t_j]$, $j=1,2,...,J$ with $0=t_0<t_1<\cdots<t_J=T$ such that

$$\rho(t,t') \quad \leq \quad C\varepsilon \quad , \quad \forall t,t'\in[t_{j-1},t_j]$$

with C denoting a finite positive constant. Furthermore, we define the partition $\mathscr{F}_{\varepsilon_j}$ as $\mathscr{F}_{\varepsilon_j}=[t_{j-1},t_j[$. After some calculations, it then follows that

$$\begin{aligned}
&E\left[\sup_{t,t'\in\mathscr{F}_{\varepsilon j}}|Z_{n_i}(t)-Z_{n_i}(t')|^2\right]\\
&\quad\leq\quad 5nh_nw_{n_i}(x;h_n)^2C^2\varepsilon^2\left\{9|\varphi_x'(1-H_x(T))|^2+2\left(\frac{\varphi_x'(1-H_x(T))}{\varphi_x'(1)}\right)^2+\frac{5}{\varphi_x'(1)^2}\right\}.
\end{aligned}$$

This implies,

$$\begin{aligned}
&\sum_{i=1}^n E\left[\sup_{t,t'\in\mathscr{F}_{\varepsilon j}}|Z_{n_i}(t)-Z_{n_i}(t')|^2\right]\\
&\quad\leq\quad 5nh_nC^2\varepsilon^2\left\{9|\varphi_x'(1-H_x(T))|^2+2\left(\frac{\varphi_x'(1-H_x(T))}{\varphi_x'(1)}\right)^2+\frac{5}{\varphi_x'(1)^2}\right\}\sum_{i=1}^n w_{n_i}(x;h_n)^2\\
&\quad\leq\quad 5nh_n||K||_2^2DC^2\varepsilon^2\left\{9|\varphi_x'(1-H_x(T))|^2+2\left(\frac{\varphi_x'(1-H_x(T))}{\varphi_x'(1)}\right)^2+\frac{5}{\varphi_x'(1)^2}\right\}
\end{aligned}$$

where D is a finite positive constant. Taking

$$C = \left(5nh_n||K||_2^2DC^2\left\{9|\varphi_x'(1-H_x(T))|^2+2\left(\frac{\varphi_x'(1-H_x(T))}{\varphi_x'(1)}\right)^2+\frac{5}{\varphi_x'(1)^2}\right\}\right)^{-1/2},$$

we see that the right hand side of the preceding inequality equals ε^2. That is, a partition $\mathscr{F}_{\varepsilon_j}$ for $\mathscr{F}=[0,T]$ constructed as described above with the appropriate choice of C satisfies (5.7). For every n, the bracketing number of this partition can be written as $N_{[]}(\varepsilon,\mathscr{F},L_2^n)=O\left(\varepsilon^{-1}\right)$. Thus, for some positive constant C', we have

$$\int_0^{\delta_n}\sqrt{\log N_{[]}(\varepsilon,\mathscr{F},L_2^n)}d\varepsilon \leq \int_0^{\delta_n}\sqrt{\log\left(\frac{C'}{\varepsilon}\right)}d\varepsilon$$

By variable transformation (i.e. substituting $u=\log(C'/\varepsilon)$), we obtain

$$\int_0^{\delta_n}\sqrt{\log\left(\frac{C'}{\varepsilon}\right)}d\varepsilon = C'\int_{\log(C'/\delta_n)}^{\infty}e^{-u}u^{1/2}du\to 0 \quad \text{as} \quad \delta_n \downarrow 0.$$

This establishes the third condition of the bracketing central limit theorem. Hence, the process $W_{xh}(\cdot)$ is asymptotically tight in $\ell^\infty[0,T]$. This together with the finite dimensional convergence establish the weak convergence of the process $(nh_n)^{1/2}\left(F_{xh}(\cdot)-F_x(\cdot)\right)$ to a Gaussian process in the space $\ell^\infty[0,T]$.

□

5.4 Numerical results

In the preceding sections of this chapter, we stated and proved some important theoretical properties of the conditional distribution function estimator under the generalized conditional Koziol-Green model. These results are asymptotic in nature and rely on the assumption that the sample size n is sufficiently large. Obviously, the following question arises: how large is "sufficiently large"? For this purpose, we set up a simulation study to investigate the performance of the estimator in finite samples in Section 5.4.1. As an illustration, we also apply the estimator to the Survival of Atlantic halibut data set in Section 5.4.2.

5.4.1 Simulation study

The finite sample performance of the generalized conditional distribution function estimator (1.25) under the generalized conditional Koziol-Green model will now be explored through a simulation study. For a desired sample size n, we consider fixed and equidistance design points $x_i = \frac{i}{n}$ $(i = 1,2,...,n)$. We assume that the survival times Y_i are independent random variables and are distributed according to

$$Y_i \sim \text{Weibull}(a_1 + a_2 x_i, b),$$

for some constants a_1, a_2, b such that $a_1 > \min(0, -a_2)$ and $b > 0$. Furthermore, we assume that the non-censoring probability for the entities depends on their design values through

$$\gamma_{x_i} \quad = \quad \frac{1}{1+\exp(a_3 + a_4 x_i)} \quad , \quad i = 1,2,...,n$$

where a_3 and a_4 are another set of constants that regulates the censoring mechanism. To generate data under the generalized conditional Koziol-Green model, we employ the conditional distribution function method (see Nelsen (2006, page 41)) as follows:

For each i $(= 1,2,...,n)$,

1. we generate two independent uniform variates $u_i \in (0,1)$ and $t_i \in (0,1)$
2. we set $v_i = c_u^{[-1]}(t_i)$, where $c_u(v) = \frac{\partial}{\partial u}\varphi_{x_i}^{[-1]}(\varphi_{x_i}(u) + \varphi_{x_i}(v))$ and $c_u^{[-1]}$ denotes the quasi-inverse of c_u
3. we obtain h_i as a solution to

$$\varphi_{x_i}(1-h_i) - \varphi_{x_i}(u_i) + \int_0^{h_i} \varphi'_{x_i}(1-w)\mathscr{C}_{x_i,01}(\gamma_{x_i}, w)dw \quad = \quad 0$$

4. we set $Y_i \quad = \quad \left(-\frac{\log(v_i)}{a_1+a_2x_i}\right)^{1/b}$ and $C_i \quad = \quad \left(-\frac{\log\left(\varphi_{x_i}^{[-1]}\left(\varphi_{x_i}(1-h_i)-\varphi_{x_i}(u_i)\right)\right)}{a_1+a_2x_i}\right)^{1/b}$
5. we set $Z_i \quad = \quad \min(Y_i, C_i)$ and $\delta_i \quad = \quad \mathbb{1}\{Y_i \leq C_i\}$

In the above algorithm, Step 2 generates a pair of uniform variables at design value $x_i \in [0,1]$ such that their joint distribution is described by an Archimedean copula with generator function φ_{x_i}. By this step, the future couple (Y_i, C_i) will satisfy

$$S_{x_i}(t_1,t_2) = P(Y_i > t_1, C_i > t_2) = \varphi_{x_i}^{[-1]}\left(\varphi_{x_i}(\bar{F}_{x_i}(t_1)) + \varphi_{x_i}\left(\bar{G}_{x_i}(t_2)\right)\right) \tag{5.8}$$

Afterwards in Step 3 and 4, we obtain the pair (Y_i, C_i) such the observable variables Z_i and δ_i also satisfies

$$H^u_{x_i}(t) = P(Z_i \leq t, \delta_i = 1) = \mathscr{C}_{x_i}(\gamma_{x_i}, H_{x_i}(t)) \tag{5.9}$$

for each $x_i \in [0,1]$.

Tables 5.1 and 5.2 summarize the simulation results for different sample sizes n, each with 10 000 replicates. The estimates are obtained at 10th, 30th, 50th, 70th and 90th percentiles Q of the marginal distribution of the survival times and correspond to prescribed time values for a given design value x. These percentiles are chosen so as to reflect the behavior of the Generalized conditional Koziol-Green estimator at various level of estimation. For the purpose of comparison, the tables also features the results under some competing estimators, namely the conditional Koziol-Green estimator of Braekers and Veraverbeke (2008) and the conditional copula graphic estimator of Braekers and Veraverbeke (2005).

The results in Tables 5.1 and 5.2 are based on the choice of parameters $a_1 = 1.5$, $a_2 = 0.5$, $b = 2$, $a_3 = -3.5$ and $a_4 = 7.5$. These parameters are chosen such that small design values are associated with smaller probability of censoring. In this way, we can easily explore the effect of censoring intensity on the estimators under consideration. For instance, estimation at design value $x = 25\%$ would reflect the behavior of the estimators on a light censored data set. Whereas estimation at design value $x = 75\%$ would provide some insight into the behavior of the estimators on a heavy censored data. For all sample sizes, Table 5.1 shows that the results based on the generalized conditional Koziol-Green estimator and conditional copula graphic estimator are close. This was expected because the copula graphic estimator is more general than the generalized conditional Koziol-Green estimator, since the former does not depend on the relationship between Z_x and δ_x. For the estimation at $x = 75\%$, we further note that the biases under

Table 5.1: Absolute biases under the generalized conditional Koziol-Green estimator F_{xh}, the conditional Koziol-Green estimator $F_{xh}^{BV(2008)}$, and the conditional copula graphic estimator $F_{xh}^{BV(2005)}$.

		$x = 25\%$			$x = 50\%$			$x = 75\%$		
	Q	F_{xh}	$F_{xh}^{BV(2008)}$	$F_{xh}^{BV(2005)}$	F_{xh}	$F_{xh}^{BV(2008)}$	$F_{xh}^{BV(2005)}$	F_{xh}	$F_{xh}^{BV(2008)}$	$F_{xh}^{BV(2005)}$
	10	0.0002	0.0091	0.0003	0.0083	0.0140	0.0065	0.0444	0.0061	0.0822
	30	0.0052	0.0169	0.0053	0.0463	0.0039	0.0635	0.0836	0.0568	0.0839
$n = 30$	50	0.0161	0.0029	0.0153	0.0636	0.0415	0.0635	0.0830	0.1480	0.0871
	70	0.0276	0.0290	0.0253	0.0615	0.0850	0.0635	0.0789	0.1841	0.0855
	90	0.0397	0.0474	0.0375	0.0586	0.0814	0.0532	0.0770	0.1333	0.0838
	10	0.0003	0.0095	0.0003	0.0047	0.0174	0.0471	0.0407	0.0107	0.0684
	30	0.0025	0.0184	0.0034	0.0359	0.0149	0.0317	0.0732	0.0493	0.0750
$n = 50$	50	0.0092	0.0085	0.0084	0.0531	0.0294	0.0470	0.0705	0.1459	0.0715
	70	0.0159	0.0169	0.0173	0.0481	0.0710	0.0479	0.0655	0.1782	0.0710
	90	0.0239	0.0309	0.0340	0.0407	0.0633	0.0411	0.0599	0.1173	0.0788
	10	0.0004	0.0086	0.0001	0.0021	0.0201	0.0039	0.0361	0.0153	0.0633
	30	0.0021	0.0175	0.0019	0.0251	0.0263	0.0196	0.0637	0.0436	0.0710
$n = 100$	50	0.0066	0.0095	0.0095	0.0412	0.0156	0.0319	0.0611	0.1497	0.0698
	70	0.0097	0.0105	0.0095	0.0360	0.0586	0.0381	0.0566	0.1783	0.0730
	90	0.0124	0.0184	0.0135	0.0246	0.0475	0.0390	0.0442	0.1042	0.0721
	10	0.0002	0.0090	0.0001	0.0016	0.0209	0.0025	0.0329	0.0179	0.0694
	30	0.0007	0.0181	0.0003	0.0201	0.0316	0.0251	0.0576	0.0395	0.0694
$n = 150$	50	0.0038	0.0117	0.0033	0.0351	0.0082	0.0412	0.0541	0.1501	0.0694
	70	0.0068	0.0074	0.0100	0.0292	0.0515	0.0385	0.0501	0.1773	0.0694
	90	0.0087	0.0145	0.0133	0.0193	0.0425	0.0155	0.0391	0.1008	0.0694
	10	0.0001	0.0088	0.0002	0.0017	0.0210	0.0019	0.0307	0.0196	0.0678
	30	0.0005	0.0178	0.0007	0.0177	0.0340	0.0200	0.0543	0.0370	0.0678
$n = 200$	50	0.0040	0.0111	0.0045	0.0317	0.0040	0.0219	0.0508	0.1507	0.0678
	70	0.0070	0.0074	0.0045	0.0267	0.0488	0.0381	0.0472	0.1770	0.0678
	90	0.0072	0.0127	0.0100	0.0159	0.0391	0.0426	0.0364	0.0988	0.0678

the generalized conditional Koziol-Green estimator are consistently smaller than their counterparts under the conditional copula graphic estimator. This can be attributed to the fact that the latter uses only the uncensored observations and due to our choice of

Table 5.2: Variances under the generalized conditional Koziol-Green estimator F_{xh}, the conditional Koziol-Green estimator $F_{xh}^{BV(2008)}$, and the conditional copula graphic estimator $F_{xh}^{BV(2005)}$.

		$x = 25\%$			$x = 50\%$			$x = 75\%$		
	Q	F_{xh}	$F_{xh}^{BV(2008)}$	$F_{xh}^{BV(2005)}$	F_{xh}	$F_{xh}^{BV(2008)}$	$F_{xh}^{BV(2005)}$	F_{xh}	$F_{xh}^{BV(2008)}$	$F_{xh}^{BV(2005)}$
	10	0.0024	0.0019	0.0039	0.0014	0.0008	0.0028	0.0009	0.0005	0.0035
	30	0.0064	0.0063	0.0093	0.0037	0.0042	0.0083	0.0075	0.0079	0.0184
$n = 30$	50	0.0112	0.0123	0.0146	0.0097	0.0109	0.0154	0.0218	0.0212	0.0374
	70	0.0143	0.0147	0.0172	0.0157	0.0151	0.0207	0.0354	0.0253	0.0436
	90	0.0092	0.0089	0.0117	0.0110	0.0101	0.0165	0.0235	0.0149	0.0372
	10	0.0016	0.0012	0.0022	0.0010	0.0006	0.0017	0.0006	0.0003	0.0020
	30	0.0042	0.0042	0.0055	0.0025	0.0028	0.0050	0.0050	0.0053	0.0118
$n = 50$	50	0.0071	0.0078	0.0087	0.0063	0.0073	0.0096	0.0154	0.0149	0.0257
	70	0.0088	0.0091	0.0101	0.0104	0.0102	0.0137	0.0239	0.0165	0.0304
	90	0.0054	0.0052	0.0064	0.0071	0.0064	0.0101	0.0148	0.0092	0.0259
	10	0.0009	0.0007	0.0011	0.0006	0.0003	0.0009	0.0003	0.0002	0.0010
	30	0.0024	0.0024	0.0029	0.0015	0.0016	0.0026	0.0032	0.0034	0.0066
$n = 100$	50	0.0041	0.0044	0.0048	0.0037	0.0045	0.0054	0.0097	0.0091	0.0153
	70	0.0050	0.0052	0.0055	0.0059	0.0059	0.0076	0.0150	0.0096	0.0187
	90	0.0028	0.0027	0.0031	0.0039	0.0034	0.0053	0.0084	0.0049	0.0172
	10	0.0006	0.0005	0.0007	0.0005	0.0002	0.0006	0.0002	0.0001	0.0007
	30	0.0017	0.0017	0.0020	0.0011	0.0012	0.0018	0.0023	0.0024	0.0047
$n = 150$	50	0.0029	0.0031	0.0033	0.0026	0.0033	0.0039	0.0073	0.0068	0.0116
	70	0.0034	0.0035	0.0037	0.0043	0.0044	0.0053	0.0111	0.0069	0.0146
	90	0.0019	0.0019	0.0021	0.0029	0.0025	0.0038	0.0061	0.0035	0.0135
	10	0.0005	0.0004	0.0005	0.0004	0.0002	0.0005	0.0002	0.0001	0.0005
	30	0.0014	0.0014	0.0016	0.0009	0.0010	0.0015	0.0018	0.0020	0.0037
$n = 200$	50	0.0023	0.0025	0.0026	0.0021	0.0026	0.0030	0.0059	0.0055	0.0093
	70	0.0027	0.0028	0.0030	0.0034	0.0034	0.0042	0.0089	0.0055	0.0119
	90	0.0015	0.0014	0.0016	0.0022	0.0019	0.0029	0.0049	0.0028	0.0116

parameters, most observations at $x = 75\%$ are censored. From Table 5.1, we also note that the biases associated with the conditional Koziol-Green estimator are in most cases larger than the corresponding ones under the generalized conditional Koziol-Green estimator. As for the classical conditional Koziol-Green estimator, Braekers and Veraver-

beke (2008) assume a relationship between Z_x and δ_x. However they take, in this case, the wrong assumption of independence.

In Table 5.2, we compare the variances associated with the simulated estimates under the three competing estimators. Since the copula graphic estimator is more general than the conditional Koziol-Green and the generalized conditional Koziol-Green estimators, we expect that the latter two estimators have smaller variances and are therefore more efficient. The results in Table 5.2 show that this is true for estimation at covariate values 25%, 50% and 75% with a sample size of at least 30. To get further insight, we repeat the simulation process for various choices of the Archimedean copula generator function φ_x and the general copula $\mathscr{C}_x$ such that (5.8) and (5.9) are respectively satisfied. However, we do not report these additional results because the conclusions are the same as above.

5.4.2 Illustration on real data: Survival of Atlantic halibut

In this section, we illustrate the generalized conditional Koziol-Green estimator on the well known Survival of Atlantic halibut data set. This data set is already introduced in Chapter 1. It is the result of a study on the size regulation of the Atlantic Halibut as one of conservation measures suggested for the trawl and long line fishery. In this section,we are particularly concerned about the survival time of a fish that was caught and handled as in the commercial fishing. In this experiment, the fish is censored by the time that it has spent in the holding tank. Some of the fishes were censored because they were removed from the holding tank within 48 hours to make space for new ones. Also, the fishes that were alive at the end of the experiment were treated as censored observations. We refer to Neilson et al. (1989) and Lange et al. (1994) for further details about this data set.

In this study, we observed that, the catch and handling are a period of great stress for the fish. In addition, the holding tank which is the fish's new environment will cause more stress. As such, one can expect a number of the fishes to die within the first few hours after they have been placed in the holding tank.

Not only the stressful environment will diminish the probability of survival for a fish,

but also, for example, an infection brought into the holding tank by a sick fish can kill the other fishes as well. Thus, it is reasonable to allege that the probability of dying from infection increases with the time spent in the holding tank. As such, the survival time Y_i of a fish depends on the time that it has spent in the holding tank C_i. Equivalently, the time spent in the holding tank has a negative influence on the survival time. Figure 5.1 is a scatter plot of survival time versus fork length of the experimental animals with a distinction between censored and uncensored observations. From the figure, it is clear that most of the censored observations occurred among fishes with fork length greater than 39 cm. This suggests possible association between censoring time and fork length of the fishes.

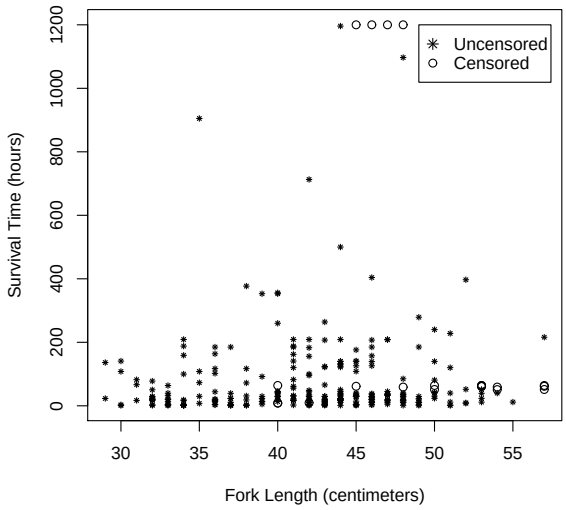

Figure 5.1: Scatter plot of fork length versus survival time with a distinction between censored and uncensored observations

A further feature of this study is that the occurrence of these censoring is a manifestation of the fishes' endurance abilities. These can inherently be attributed to several factors such as the time of the catch and the prevailing atmospheric temperature or wind, among others. As a result, we suspect that censoring time could be additionally infor-

mative to the survival time through its distribution function. As such, the generalized conditional Koziol-Green estimator could be the outstanding candidate to estimate the survival distribution of the time until death, provided relation (5.9) holds. In this practical data illustration, we verify condition (5.9) by looking at the empirical counterpart and investigate whether, at a desired fork length x

$$H^u_{xh}(t) \quad = \quad \mathscr{C}_x\left(\gamma_{xh}, H_{xh}(t)\right) \tag{5.10}$$

nearly holds for all $t \geq 0$, where

$$\begin{aligned} H^u_{xh}(t) &= \sum_{i=1}^{n} w_{ni}(x,h_n)\mathbb{1}\{z_i \leq t, d_i = 1\} \\ H_{xh}(t) &= \sum_{i=1}^{n} w_{ni}(x,h_n)\mathbb{1}\{z_i \leq t\} \\ \gamma_{xh} &= \sum_{i=1}^{n} w_{ni}(x,h_n)\mathbb{1}\{d_i = 1\} \end{aligned}$$

with z_i and d_i denoting respectively, the observed time and censoring indicator at fork length x_i $(i = 1,2,3,...,n)$ and $w_{ni}(x,h_n)$ is the corresponding weight at fork length x (compare with definitions (1.12) and (1.13) in Chapter 1).

Relation (5.10) readily suggests an informal procedure to investigate the relationship between the observed time and censoring indicator. In other words, a plausible function to describe the relationship between the observed variables at a given covariate value x needs to give the best approximation to $H^u_{xh}(t)$, uniformly over $t \geq 0$. In Figure 5.2, we present a visual test for the copula function to describe the relationship between the observed time and censoring indicator as defined in (5.10). In particular, we consider estimation at fork lengths $(= 32, 53)$ and compare Fréchet-Hoeffding lower bound (F-H lower), Frechet-Hoffding upper bound (F-H upper), Product, Gumbel bivariate logistic, Plackett (with parameter $\theta = 10$), Clayton (with parameter $\theta = 3$) and Frank (with parameter $\theta = 8$) copulas to the empirical quantity $H^u_{xh}\left(H^{-1}_{xh}(p)\right)$, where $H^{-1}_{xh}(p) = \inf\{t : H_{xh}(t) > p\}$ is the quantile function of $H_{xh}(t)$. These copula functions are given in Nelsen (2006). For convenience, we list them here respectively in a general

way as follows:

$$\begin{aligned}
\mathscr{C}(u,v) &= \max(u+v-1,0)\\
\mathscr{C}(u,v) &= \min(u,v)\\
\mathscr{C}(u,v) &= uv\\
\mathscr{C}(u,v) &= \frac{uv}{u+v-uv}\\
\mathscr{C}(u,v) &= \frac{1+(\theta-1)(u+v)-\sqrt{[1+(\theta-1)(u+v)]^2-4uv\theta(\theta-1)}}{2(\theta-1)}\\
\mathscr{C}(u,v) &= \left[\max\left(u^{-\theta}+v^{-\theta}-1,0\right)\right]^{-1/\theta}\\
\mathscr{C}(u,v) &= -\frac{1}{\theta}\log\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)
\end{aligned}$$

Irrespective of the bandwidths (i.e. $h_n = 20, 40$), we notice from the first row of Figure 5.2 that the Gumbel logistic copula gives the best approximation to the empirical quantity. As a result, we consider it as the most plausible candidate to describe the relationship between observed times and the censoring indicators in this data set. However, this observation is not clear in the second row of Figure 5.2, since the various copula function approximations almost coincide with the empirical quantity $H^u_{xh}\left(H^{-1}_{xh}(p)\right)$. This is expected because there is no censoring at fork length $= 32$, as can be seen in Figure 5.1. As a consequent, $\gamma_{xh} \approx 1$, $H^u_{xh}(t) \approx H_{xh}(t)$ for all time $t \geq 0$ and

$$H^u_{xh}(t) = \mathscr{C}_x(\gamma_{xh}, H_{xh}(t)) \approx H_{xh}(t),$$

for all copula functions under consideration. Using the Gumbel logistic copula, we present in Figure 5.3, the generalized conditional Koziol-Green estimate of the distribution function of time until death in the holding tank. In this estimation, we assume the Frank copula with generator function

$$\varphi_x(u) = -\log\left(\frac{e^{(x-20)u}-1}{e^{x-20}-1}\right)$$

for the association structure of the survival time and censoring time. This choice has also been considered by Braekers and Veraverbeke (2005) for the conditional copula graphic estimator. It allows the dependence structure of the survival time and censoring time to

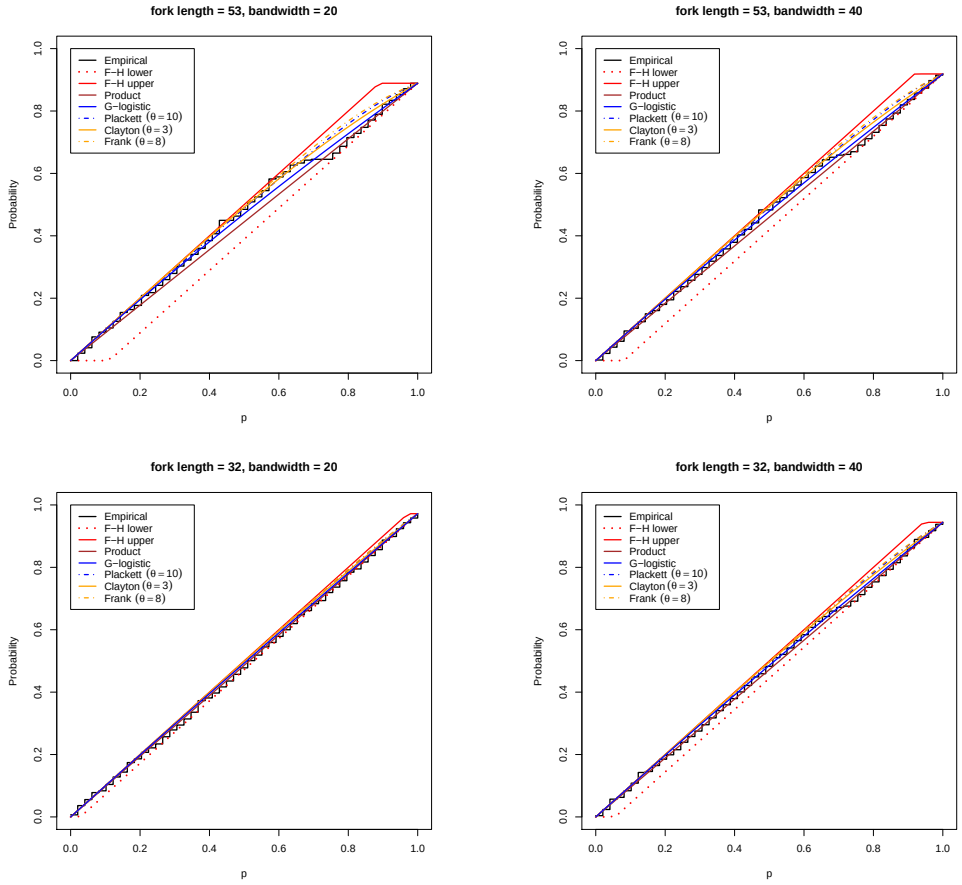

Figure 5.2: Graphical test for the copula function to describe the relationship between observed time and censoring indicator.

also depend on the fork length x and gives a stronger discordance association for larger fishes, given that all the fishes are kept in the holding tank for the same amount of time. Equivalently, this means that the survival probability of larger fishes will be smaller than the survival probability of the smaller fishes.

For the purpose of comparison, Figure 5.3 also features the conditional copula graphic and the conditional Koziol-Green estimates of the survival distribution. Due to the large proportion of uncensored observations in this data set, we observe from Figure 5.3 that the three survival distribution estimates are close to each other. This is because, the

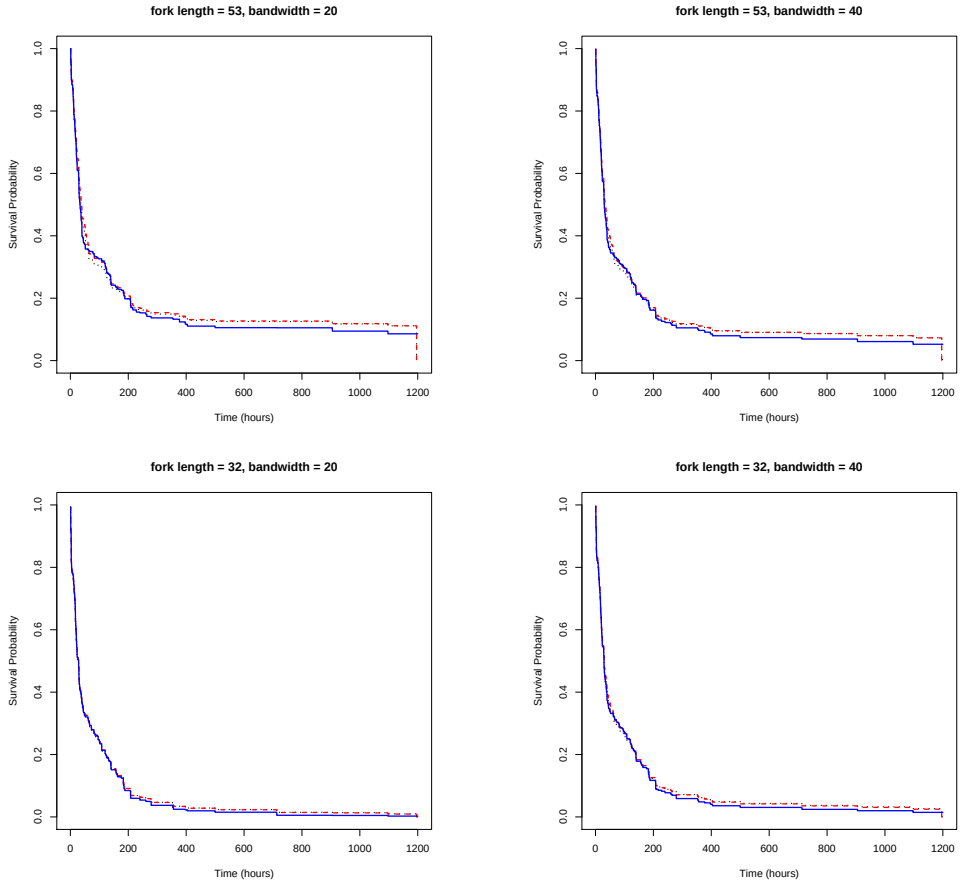

Figure 5.3: Estimates of the survival probabilities of Atlantic halibuts under the generalized conditional Koziol-Green estimator (black dotted line), the conditional copula graphic estimator (blue continuous line) and the conditional Koziol-Green estimator (red dashed line)

performance of the generalized conditional estimator over the other two tends to increase with decreasing proportion of uncensored observations. This is more obvious in the second row of Figure 5.3 where all three competing estimators coincide, due to the high probability of non-censoring at fork length $= 32$. In general, it is obvious that the Gumbel bivariate copula captures the possible relationship between the observed times and the censoring indicator. In addition, Figure 5.3 shows that the effect of the bandwidth choice on the conclusions under these three estimators is negligible.

Nonetheless, it might be appropriate to point out that the graphical test presented in Figure 5.1 is for illustrative purpose only. A formal test of copula functions based on the observable variables Z_i and δ_i is possible and follows the lines of Chapter 3. On the contrary, it is not possible to perform a formal test on the choice of Archimedean generator function $\varphi_x(u)$ to model the dependence structure of the survival time and censoring time, since neither of these random variables is fully observable. By following the work of Braekers and Veraverbeke (2005), we can however, conduct a thorough sensitivity analysis on the choice of φ_x. We do not dwell on these any further in this thesis, but they may be possible areas for future exploration.

6

Possible future research

In Chapter 1, we introduced the generalized conditional Koziol-Green estimator (1.25). This estimator is nonparametric and depends on some general copula function $\mathscr{C}_x$ that is assumed to be known *a priori*. Nevertheless, there are cases where we may be interested in a certain copula family. For example, when the Plackett copula is used on the observable variables Z_i and δ_i $(i = 1,2,...,n)$, then its parameter is the constant odds ratio for the conditional sub-distributions of the observed variables. In such cases, it might be appropriate to adopt a semiparametric form of (1.25). To achieve this, we parameterize the generalized conditional Koziol-Green model (1.18) and assume instead that

$$\bar{G}_x(t) \quad = \quad \nu_x(\theta, \bar{F}_x(t)) \tag{6.1}$$

with, for every value of θ, $\nu_x(\theta,\omega)$ a non-decreasing function of $\omega \in [0,1]$, $\nu_x(\theta,0) = 0$ and $\nu_x(\theta,1) = 1$. Similar to the nonparametric derivation (see in Section 1.2.2), we find this function $\nu_x(\cdot,\cdot)$ such that the sub-distribution at $x \in [0,1]$ of the uncensored observations satisfies

$$H_x^u(t) \quad = \quad P(Z_x \leq t, \delta_x) = \mathscr{C}_x(\theta; \gamma_x, H_x(t)) \tag{6.2}$$

where $\{\mathscr{C}_x(\theta;\cdot,\cdot) : \theta \in \Theta\}$ is the desired copula family depending on some unknown parameter θ from a compact parameter space $\Theta \in \mathbb{R}^d$, with $d \in \mathbb{N}$. By following the lines leading (1.25), we easily obtain the semiparametric estimator

$$\bar{F}_{xh}(t) \quad = \quad \varphi_x^{[-1]}\left(-\int_0^{H_{xh}(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}\left(\hat{\theta}; \gamma_{xh}, w\right) dw\right) \tag{6.3}$$

in the generalized conditional Koziol-Green model (6.1) under dependent censoring. For this semiparametric estimator, γ_{xh} and $H_{xh}(t)$ are the nonparametric Stone (1977) type estimators of γ_x and $H_x(t)$. To obtain the estimator $\hat{\theta}$, we propose a likelihood based technique. More specifically, we note from (6.2) that the likelihood contribution of the ith data point is

$$\mathscr{L}_i(\theta) = \begin{cases} \mathscr{C}_{x_i,01}(\theta; \gamma_{x_i}, H_{x_i}(z_i)) & \text{if } d_i = 1 \\ 1 - \mathscr{C}_{x_i,01}(\theta; \gamma_{x_i}, H_{x_i}(z_i)) & \text{if } d_i = 0 \end{cases}.$$

Obviously, this leads to the likelihood

$$\prod_{i=1}^n \mathscr{C}_{x_i,01}(\theta; \gamma_{x_i}, H_{x_i}(z_i))^{d_i} (1 - \mathscr{C}_{x_i,01}(\theta; \gamma_{x_i}, H_{x_i}(z_i)))^{1-d_i}$$

where z_i and d_i $(i = 1,2,...,n)$ are the observed times and censoring indicators respectively. However, we note in this likelihood that γ_{x_i} and $H_{x_i}(\cdot)$ are unknown. Replacing them by their respective Stone (1977) type counterparts (see Chapter 1), subsequently yields the pseudo-likelihood

$$\prod_{i=1}^n \mathscr{C}_{x_i,01}(\theta; \gamma_{x_i h}, H_{x_i h}(z_i))^{d_i} (1 - \mathscr{C}_{x_i,01}(\theta; \gamma_{x_i h}, H_{x_i h}(z_i)))^{1-d_i}.$$

At a design value $x \in [0,1]$, we subsequently obtain $\hat{\theta}$ as a solution to the weighted score equation

$$\sum_{i=1}^n w_{n_i}(x, h_n) \left\{ d_i \frac{\mathscr{C}'_{x_i,01}(\gamma_{x_i h}, H_{x_i h}(z_i))}{\mathscr{C}_{x_i,01}(\gamma_{x_i h}, H_{x_i h}(z_i))} - (1-d_i) \frac{\mathscr{C}'_{x_i,01}(\gamma_{x_i h}, H_{x_i h}(z_i))}{1 - \mathscr{C}_{x_i,01}(\gamma_{x_i h}, H_{x_i h}(z_i))} \right\} = 0$$

where $w_{n_i}(x,h_n)$ is the Gasser-Müller weight defined in Chapter 1 and

$$\mathscr{C}'_{x_i,01}(\theta;u.v) = \left(\frac{\partial}{\partial\theta_1}\mathscr{C}_{x_i,01}(\theta;u.v),...,\frac{\partial}{\partial\theta_q}\mathscr{C}_{x_i,01}(\theta;u.v)\right)^T$$

is a vector of partial derivatives for each component of θ. In the weighted score equation above, each pseudo log-likelihood contribution is multiplied by the weight. In this way, those observations x_i close to x have a higher impact in estimating θ.

6.1 Theoretical properties

As in the previous chapters, one important step towards the proposed semiparametric estimator (6.3) is to show its consistency as an estimator of the true survival distribution function $\overline{F}_x$ at $x \in [0,1]$ as well as the weak convergence of the corresponding empirical process to an appropriate Gaussian process with some variance covariance function. To carry out these, we first need to establish the consistency and normality of the preliminary estimators γ_{xh}, H_{xh} and $\hat{\theta}$. For γ_{xh} and H_{xh}, these results has already been established and can be found in the literature. See for example Van Keilegom and Veraverbeke (1997a) and Braekers and Veraverbeke (2001), among others. For $\hat{\theta}$, we could adapt the results of Newey (1994) and Chen et al. (2003), who gave primitive conditions under which a semiparametric estimator that depends on some preliminary nonparametric estimators is consistent and asymptotically normal.

Once the important asymptotic results of the preliminary estimators are established, we can proceed in parallel with Chapter 5 and ascertain the desired theoretical properties of the semiparametric estimator (6.3). Furthermore, it might be appropriate to determine the validity of the generalized semiparametric conditional Koziol-Green estimator in practical applications. Similar to Chapter 3, this will reduce to testing for the null hypothesis

$$H_0: \quad H_x^u(t) - \mathscr{C}_x(\theta_o;\gamma_x,H_x(t)) = 0 \ \text{ for all } \ t \geq 0,$$

due to the infeasibility of a formal test to ascertain the dependence structure that governs the joint distribution of the survival time and censoring time. For the alternative

hypothesis, we may allow for any deviation of the conditional sub-distribution of the uncensored observations from the γ_x-section of the general copula function $\mathscr{C}_x(\theta;.,.)$. To establish the necessary conditions for the validity of the estimator (6.3), we can further mimic the theoretical development of the testing procedure presented in Chapter 3. In line with Chapter 3, it might also be convenient to consider a bootstrap approximation of the testing procedure.

Bibliography

1. Abdushukurov, A., A. (1987). Nonparametric estimation in the proportional hazards model of random censorship. *Akad. Nauk. Uz Tashkent VINITI NO.*, 3448–V.

2. Aerts, M., P. Janssen, and N. Veraverbeke (1994). Bootstrapping regression quantiles. *Journal of Nonparametric Statistics 4*, 1–20.

3. Andersen, K. P., Ø. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

4. Araujo, A. and E. Giné (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables*. New York: Wiley.

5. Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.

6. Braekers, R. and A. Gaddah (2007). A Koziol-Green estimator for the conditional distribution function under dependent censoring. In Gomes M. I., Pestana D., Silva P. (Eds). *The 56th Session of the International Statistical Institute - Proceedings*.

7. Braekers, R. and A. Gaddah (2011). Flexible modeling in the Koziol-Green model by a copula function. *Communications in Statistics - Theory and Methods*, In–press.

8. Braekers, R. and N. Veraverbeke (2001). The partial Koziol-Green model with covariates. *Journal of Statistical Planning and Inference 92*, 55–71.

9. Braekers, R. and N. Veraverbeke (2003). Testing for the partial Koziol-Green model with covariates. *Journal of Statistical Planning and Inference 115*, 181–192.

10. Braekers, R. and N. Veraverbeke (2005). Bootstrapping the conditional survival function estimator in the partial Koziol-Green model. *Journal of Nonparametric Statistics 17*,

299–318.

11. Braekers, R. and N. Veraverbeke (2008). The conditional Koziol-Green model under dependent censoring. *Statistics and Probability Letters 78*, 927–937.

12. Cao, R., I. López-De-Ullibarri, P. Janssen, and N. Veraverbeke (2005). Presmoothed Kaplan-Meier and Nelsen-Aalen estimators. *Nonparametric Statistics 17*, 31–56.

13. Chen, X., O. Linton, and I. Van Keilegom (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica 71*, 1591–1608.

14. Cheng, P. and G. Lin (1987). Maximum likelihood estimation of a survival function under the Koziol-Green proportional hazards model. *Statistics and Probability Letters 5*, 75–80.

15. Csörgó, S. (1988). Testing for the proportional hazards model of random censorship. *Proceedings of the 4th Praque Symposium on Asymptotic Statistics, Praque 1*, 87–92,.

16. Dikta, G. (1998). On semiparametric random censorship models. *Journal of Statistical Planning and Inference 66*, 253–279.

17. Foran, J. (1991). *Fundamentals of Real Analysis*. New York: Marcel Dekker.

18. Gaddah, A. and R. Braekers (2009). Weak convergence for the conditional distribution function in a Koziol-Green model under dependent censoring. *Journal of Statistical Planning and Inference 139*, 930–943.

19. Gaddah, A. and R. Braekers (2010a). An extension of the Koziol-Green model under dependent censoring. *Journal of Nonparametric Statistics (Accepted)*.

20. Gaddah, A. and R. Braekers (2010b). Testing under the extended Koziol-Green model. In Durante, F., Hardle, W., Jaworski, P., Rychlik, T. (Eds). *Workshop on Copula Theory and its Applications, Lecture Notes in Statistics - Proceedings. Springer, Berlin/Heidelberg 98*, 279–288.

21. Genest, C. and J. Nešlehová (2007). A primer on copulas for count data. *Austin Bulletin 37*, 475–515.

22. Hollander, M. and E. Pĕna (1989). Famillies of confidence bands for the survival function under the general random censorship model and the Koziol-Green model. *Canadian Journal of Statistics 17*, 59–74.

23. Hosmer, W. D. and S. Lemeshow (1999). *Applied Survival Analysis: Regression Modelling of Time to Event Data*. New York: Wiley.

24. Kaplan, E. L. and P. Meier (1958). Non-parametric estimation from incomplete obser-

vations. *Journal of American Statistical Association 53*, 457–481.

25. Klement, P. E., A. Kolesárová, R. Mesiar, and C. Sempi (2007). Copulas constructed from horizontal sections. *Communications in Statistics- Theory and Methods 36*, 2901–2911.

26. Kochar, C. S. and F. Proschan (1991). Independence of time and cause of failure in the multiple dependent competing risks model. *Statistica Senica 1*, 295–299.

27. Koziol, J. A. and S. B. Green (1976). A Cramér-von Mises statistic for randomly censored data. *Biometrika 63*, 465–474.

28. Lange, N., L. Ryan, L. Billard, D. Brillinger, L. Conquest, and J. Greenhouse (1994). *Case Studies in Biometry*. New York: Wiley.

29. Miller, G. R. (1981). *Survival Analysis*. USA: Wiley.

30. Neilson, J., K. Waiwood, and S. Smith (1989). Survival of Atlantic halibut (Hippoglossus hippoglossus) caught by longline and otter trawl gear. *Canadian Journal of Fisheries and Aquatic Sciences 46*, 887–897.

31. Nelsen, R. B. (2006). *An Introduction to Copulas*. New York: Springer-Verlag.

32. Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica 62*, 1349–1382.

33. Rao, C. (1973). *Linear Statistical Inference and its Applications*. New York: Wiley.

34. Rivest, L. and M. T. Wells (2001). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis 79*, 138–155.

35. Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

36. Sethuraman, J. (1965). On a characterization of the three limiting types of the extreme. *Sankhya. Series A 27*, 357–364.

37. Stone, C. J. (1977). Consistent non-parametric regression. *The Annals of Statistics 5*, 595–645.

38. Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences of the United States of America 72*, 20–22.

39. van der Vaart, A. and J. Wellner (2000). *Weak Convergence and Empirical Processes*. New York: Springer.

40. van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University

Press.

41. Van Keilegom, I. and N. Veraverbeke (1997a). Estimation and bootstrap with censored data in fixed design nonparametric regression. *Annals of the Institute of Statistical Mathematics 49*, 467–491.

42. Van Keilegom, I. and N. Veraverbeke (1997b). Weak convergence of the bootstrapped conditional Kaplan-Meier process and its quantiles process. *Communications in Statistics: Theory and Methods 26*, 853–869.

43. Veraverbeke, N. and C. Cadarso-Suárez (2000). Estimation of the conditional distribution in a conditional Koziol-Green model. *Test 9*, 97–122.

44. Zheng, M. and J. Klein (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika 82*, 127–138.

Samenvatting

In hedendaagse wetenschappelijke onderzoeken, ontmoeten wij vaak studies waarvan de primaire interesse gericht is op niet-negatieve reactievariabelen. Een specifiek voorbeeld hiervan zijn de overlevingsstudies waarbij de reactievariabele de tijd tot een bepaalde gebeurtenis is. Dit soort studies worden toegepast in diverse onderzoeksgebieden. In techniek bijvoorbeeld, is de tijd tot het falen van een machinecomponent belangrijk. Bij de sociale wetenschappen, kijkt men naar de duur van stakingen, de duur van werkloosheid of de duur van huwelijken in sommige maatschappijen. In een medische context heeft men overlevingsstudies wanneer men de tijd tot hervallen na een kankertumor, de tijd tot herstel na een chirurgische operatie of de levensduur van sommige biologische organismen onderzoekt. Bij verschillende overlevingsstudies staat de responsvariabele ”tijd” niet voor de letterlijke tijd. Bijvoorbeeld, in kwaliteitscontrole of betrouwbaarheid in de productie is men meestal geïnteresseerd in de kracht die nodig is om een onderdeel onbruikbaar te maken. Terwijl in economie, stelt dit het bedrag voor dat door een verzekeringsmaatschappij wordt betaald in het geval van schade.

Door verschillende praktische redenen kunnen we de response variabelen niet volledig waarnemen voor elke studieobject. Bij enkele studieobjecten nemen wij hun exacte responstijd waar, terwijl voor anderen objecten is slechts gedeeltelijke informatie voor de responstijd beschikbaar. Eén bron van gedeeltelijke informatie is censurering. Bijvoorbeeld in een medische studie waar de overlevingstijd door een hartkwaal belangrijk is. In

dit geval, het is mogelijk dat sommige patiënten sterven aan andere ziekten zodat hun exacte overlevingstijd niet kan worden waargenomen. Namelijk wanneer een patiënt sterft zonder een hartkwaal, is de enige informatie over de overlevingstijd van een hartkwaal dat dit groter is dan de waargenomen overlevingstijd. We noemen deze overlevingstijd rechts gecensureerd.

Ondanks dat censurering een integraal deel vormt van overlevingsstudies, heeft dit gevolgen voor het trekken van conclusies in dergelijke studies. Voorts moeten we niet-verifieerbare veronderstellingen maken over de associatie tussen de respons (overlevings) variabele en de censureringsvariabele. Onder de veronderstelling dat de overlevings en censureringsvariabele onafhankelijk zijn, vormt de Kaplan en Meier (1958) schatter een inferenciële basis voor de verdeling van de overlevingstijd.

In sommige studies zien we dat de censureringsvariabele informatief is voor de overlevingsveranderlijke Y door zijn distributiefunctie. Om deze informatieve censurering te behandelen, stelde Koziol en Green (1976) een submodel voor waarbij de distributiefunctie F van de overlevingsveranderlijke Y en de distributiefunctie G van de censureringsveranderlijke voldeden aan de volgende relatie

$$1-G(t) \quad = \quad (1-F(t))^{\beta} \quad , \ \beta > 0.$$

Volgens Kochar en Proschan (1991) , kunnen we gemakkelijk aantonen dat de voorafgaande karakterisering van dit submodel gelijkwaardig is aan het feit dat de waarneembare variabelen $Z = \min(Y,C)$ en $\delta = \mathbb{1}\{Y \leq C\}$ onafhankelijk zijn. Gebaseerd op deze extra veronderstelling en de onafhankelijkheid van Y en C, hebben Abdushukurov (1987) en Cheng en Lin (1987) gevonden dat

$$\bar{F}^{ACL}(t) \quad = \quad (1-H(t))^{\gamma},$$

waarbij γ het percentage van ongecensureerde observaties is en $H(\cdot)$ de distributiefunctie van Z is. Door het vervangen van γ en $H(\cdot)$ door respectievelijke empirisch schatters γ_n en $H_n(\cdot)$, de auteurs verkreeg de volgende schatter van de overlevingstijddistributie

$$\bar{F}_n^{ACL}(t) \quad = \quad (1-H_n(t))^{\gamma_n}.$$

Zij bestudeerden de asymptotische eigenschappen van de schatter en toonden zijn superioriteit over Kaplan and Meier (1958) schatter in termen van asymptotische efficiency.

Volgens Csörgó (1988) is de praktische toepasbaarheid van $\bar{F}_n^{ACL}(\cdot)$ beperkt aangezien het kleine aantal data sets waarvoor de onafhankelijkheidsveronderstelling op Z en δ houdt. In Hoofdstuk 1 van de thesis, introduceerden wij een uitbreiding van $\bar{F}_n^{ACL}(\cdot)$ waarbij we de onderliggende onderstellingen generaliseren. Namelijk enerzijds hebben we de associatie tussen Y en C, en anderzijds is er een verband tussen F en G. Voor de eerste veronderstelling nemen we aan zoals in Rivest en Wells (2001) dat de gezamenlijke overlevingsdistributie van Y en C voldoet aan

$$S(t_1,t_2) \quad = \quad P(Y > t_1, C > t_2) = \varphi^{[-1]}\left(\varphi(\bar{F}(t_1)) + \varphi(\bar{G}(t_2))\right),$$

waarbij $\bar{F}(t) = 1 - F(t)$ en $\bar{G}(t) = 1 - G(t)$ de respectievelijke overlevingsdistributies zijn van Y en C. De functie $\varphi : [0,1] \to [0,\infty]$ is een bekende generator van de Archimedische copula functie. We noteren $\varphi^{[-1]}$ voor de pseudo-inverse van deze generator zoals weergegeven in Nelsen (2006). Voor de tweede onderstelling veralgemenen wij het verband tussen F en G indirect door een andere copula functie $\mathscr{C}$ op de waarneembare variabelen Z en γ zodat de sub-distributie van de ongecensureerde observaties gegeven wordt door

$$H^u(t) \quad = \quad P(\delta = 1, Z \leq t) = \mathscr{C}(\gamma, H(t)),$$

waarbij $\gamma = P(\delta = 1)$ het verwachte aandeel ongecensuurde observaties is en $H(t) = P(Z \leq t)$ is de distributie van het waargenomen overlevingstijd. Wij bestuderen de asymptotische eigenschappen van de uitgebreide schatter in Hoofdstuk 2 en testen zijn toepasselijkheid door een goodness-of-fit-procedure in Hoofdstuk 3.

In Hoofdstukken 4 en 5 beschouwen wij de situatie waarbij enkele extra gemeten variabelen (covariaten) beschikbaar zijn. Deze covariaten zijn in de meeste voorbeelden niet van primair belang voor de onderzoeker, maar ze hebben het potentieel om de distributie van de overlevingstijd te beïnvloeden. Als voorbeeld, denken wij aan een studie die inzicht probeert te geven in de distributie van de lengte van een verblijf voor patiënten in een ziekenhuisopname. Hierbij is het duidelijk dat de distributie van de overlevingstijd (duur van het ziekenhuisverblijf) door de leeftijd en/of een medische conditie (ernstigheidsgraad van de ziekte) van de patiënt bij opname kan worden beïnvloed. Voor een meer technische voorstelling van dit probleem, veronderstellen we dat $Y_{x_1}, Y_{x_2}, , Y_{x_n}$ onafhankelijke overlevingstijden zijn op vaste designpunten $x_1 < x_2 < \cdots < x_n$. Elke

Y_{x_i} is geassocieerd met een censurerende veranderlijke C_{x_i} . Bij elk ontwerppunt x_i, zijn de waarneembare variabelen $Z_{x_i} = \min(Y_{x_i}, C_{x_i})$ en $\delta_{x_i} = \mathbb{1}\{Y_{x_i} \leq C_{x_i}\}$. Wij behandelen deze regressie setting op twee manieren. Eerst, veronderstellen wij dat bij een bepaalde covariaat waarde $x \in [0,1]$, de distributiefunctie F_x van de overlevingstijd Y_x bij x en de distributiefunctie G_x van de censuringsveranderlijke C_x bij x voldoet aan de voorwaardelijke Koziol-Green karakterisering

$$1 - G_x(t) \quad = \quad (1 - F_x(t))^{\beta_x},$$

waarbij $\beta_x > 0$ slechts afhangt van x. Gelijkaardig als in de situatie zonder covariaten, veronderstellen we dat de mogelijke afhankelijkheid tussen Y_x en C_x gegeven is een Archimedische copula functie die voldoet aan

$$S_x(t_1,t_2) \quad = \quad P(Y_x > t_1, C_x > t_2) = \varphi_x^{[-1]}\left(\varphi_x(\bar{F}_x(t_1)) + \varphi_x(\bar{G}_x(t_2))\right).$$

Deze assumptie samen met de voorwaardelijke Koziol-Green karakterisering leidt tot het voorwaardelijke Koziol-Green model

$$\bar{F}_x^{BV}(t) \quad = \quad \varphi_x^{[-1]}\left(\gamma_x \varphi_x(\bar{H}_x(t))\right),$$

voor de voorwaardelijke overlevingsdistributie van de overlevingstijd onder afhankelijke censurering waarbij $\gamma_x = P(\delta_x = 1)$ en $H_x(t) = P(Z_x \leq t)$. In dit model krijgen we de schatter

$$\bar{F}_{xh}^{BV}(t) \quad = \quad \varphi_x^{[-1]}\left(\gamma_{xh} \varphi_x(\bar{H}_{xh}(t))\right),$$

met γ_{xh} het gewogen percentage ongecensureerde observaties en $H_{xh}(\cdot)$ de gewogen empirische distributie van de waargenomen overlevingstijd. De voorafgaande schatter werd voorgesteld door Braekers en Veraverbeke (2008). De auteurs toonden zijn consistentie en asymptotische normaliteit aan. In Hoofdstuk 4, complementeren wij hun resultaat met de zwakke convergentie van het bijbehorende proces. Gebruikmakend van dit recentere resultaat toonden wij de asymptotische efficiency aan van $\bar{F}_{xh}^{BV}$ over de copula-graphic schatter van Braekers en Veraverbeke (2005) . In hetzelfde hoofdstuk, ontwikkelden wij een betrouwbaarheidsband voor $\bar{F}_{xh}^{BV}$ en illustreren deze op een praktische data set -Worcester Heart Attack Study.

Overeenkomstig het scenario zonder covariates, veralgemenen wij de onafhankelijkheidsbeperking op Z_x en δ_x en verkrijgen, na wat algebra, de algemene voorwaardelijke Koziol-Green schatter

$$\bar{F}_{xh}(t) \quad = \quad \varphi_x^{[-1]}\left(-\int_0^{H_{xh}(t)} \varphi_x'(1-w)\mathscr{C}_{x,01}(\gamma_{xh},w)dw\right),$$

met γ_{xh} en H_{xh} zoals vroeger bepaald. In Hoofdstuk 5, bestudeerden wij $\bar{F}_{xh}$ en maakten zijn consistentie en zwakke convergentieresultaten duidelijk. Verder, onderzoeken wij zijn eindige steekproefeigenschappen via een simulatiestudie en illustreren het op de "Atlantic halibut data set".