Bernstein-based estimation of the cross ratio function

Reference:
Abrams Steven, Sercik Omer, Veraverbeke Noel.- Bernstein-based estimation of the cross ratio function
Statistics: a journal of theoretical and applied statistics - ISSN 1029-4910 - 58:1(2024), p. 230-246
Full text (Publisher's DOI): https://doi.org/10.1080/02331888.2024.2320924
To cite this reference: https://hdl.handle.net/10067/2047660151162165141

ARTICLE TEMPLATE

Bernstein-based estimation of the cross ratio function

Steven Abrams[a,b], Ömer Sercik[a] and and Noël Veraverbeke[a,c]

[a] Data Science Institute, Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium; [b] Global Health Institute, Department of Family Medicine and Population Health, University of Antwerp, Wilrijk, Belgium; [c] School of Computer Science, Statistics and Mathematics, North-West University, Potchefstroom Campus, Potchefstroom, South Africa

ARTICLE HISTORY
Compiled January 24, 2024

ABSTRACT
Local association measures provide useful insights in time-varying changes in association, especially between time-to-event variables. Such local dependence between two correlated random variables can be measured using the cross ratio function, introduced by Clayton (1978). The cross ratio function is defined as the ratio of conditional hazard functions which have been estimated using Bernstein polynomials before. Alternatively, the cross ratio function can be expressed in terms of (derivatives of) the joint survival function of the two random variables. In this paper, we discuss an alternative Bernstein-based plug-in estimator of the cross ratio function in which each of the ingredients is estimated separately. Next to asymptotic normality of the nonparametric estimator, a simulation study is used to assess its finite-sample performance. Finally, the novel estimator is applied to a real-life data application.

KEYWORDS
Bernstein polynomials; copula functions; local association; time-to-event data; survival functions.

1. Introduction

The cross ratio function (CRF) is a measure for the association between two non-negative time-to-event random variables T_1 and T_2 [1], defined as

$$\theta(t_1, t_2) = \frac{\lambda(t_1|T_2 = t_2)}{\lambda(t_1|T_2 > t_2)} \tag{1}$$

for event times t_1 and t_2 being realisations of T_1 and T_2, respectively, and with $\lambda(\cdot|T_2 = t_2)$ and $\lambda(\cdot|T_2 > t_2)$ representing the conditional hazard rate functions for T_1 given $T_2 = t_2$ and $T_2 > t_2$, respectively. Independence between T_1 and T_2 corresponds to $\theta(t_1, t_2) \equiv 1$ and positive association corresponds to $\theta(t_1, t_2) > 1$. See also Oakes [2–4] for a further discussion with regard to the cross ratio function. Several authors have discussed the use of the cross ratio function to estimate local dependence between T_1 and T_2 either in a parametric [5], semi-parametric or non-parametric way [6].

CONTACT S. Abrams. Email: steven.abrams@uhasselt.be

Alternatively, the CRF can be expressed in terms of the derivatives of the joint survival function of the time-to-event random variables T_1 and T_2 as follows:

$$\theta(t_1, t_2) = \frac{f(t_1, t_2)S(t_1, t_2)}{[\partial S(t_1, t_2)/\partial t_1]\,[\partial S(t_1, t_2)/\partial t_2]}. \tag{2}$$

In this paper, we focus on the use of an alternative non-parametric estimator for the CRF, inspired by earlier work by [6]. More specifically, we focus on the Bernstein-based estimation of the different components of the CRF in equation (2). The estimator in [6] uses Bernstein estimators for the conditional hazard functions in the numerator and denominator of (1). However, this requires an additional kernel smoothing approach leading to a bandwidth to be selected together with the Bernstein order. Therefore, the novel plug-in estimator proposed in this paper is more straightforward. It is known that Bernstein estimation provides a good order of the bias, uniformly on the unit square, and its good performance has been demonstrated in the past in a series of papers by Janssen, Swanepoel and Veraverbeke (see [7–9]) when estimating the copula function and copula derivatives.

More specifically, the paper is organized as follows. First, we introduce the notation and terminology used throughout the paper as well as the novel estimator in Section 2. We study asymptotic properties of the new estimator in Section 3 and investigate finite-sample performance using a detailed simulation study in Section 4. The use of the estimator is illustrated using a real-life data application in Section 5. Finally, we end with a discussion on future extensions and avenues for further research in Section 6.

2. Estimation of the cross-ratio function

Let (T_1, T_2) represent a random vector of non-negative time-to-event random variables with joint and marginal survival functions

$$\begin{aligned} S(t_1, t_2) &= P(T_1 > t_1, T_2 > t_2), \\ S_1(t_1) &= P(T_1 > t_1) \\ S_2(t_2) &= P(T_2 > t_2), \end{aligned}$$

and corresponding densities $f(t_1, t_2)$, $f_1(t_1)$ and $f_2(t_2)$. Based on Sklar's theorem, the joint survival function $S(t_1, t_2)$ can be written in terms of the survival copula $C(\cdot, \cdot)$ and the marginal survival functions as follows:

$$S(t_1, t_2) = C\left[S_1(t_1), S_2(t_2)\right].$$

The cross-ratio function is defined as in equation (1) where

$$\begin{aligned} \lambda(t_1|T_2 = t_2) &= \lim_{\Delta\to 0^+} \frac{1}{\Delta} P(t_1 < T_1 < t_1 + \Delta \mid T_1 > t_1, T_2 = t_2) \\ \lambda(t_1|T_2 > t_2) &= \lim_{\Delta\to 0^+} \frac{1}{\Delta} P(t_1 < T_1 < t_1 + \Delta \mid T_1 > t_1, T_2 > t_2). \end{aligned}$$

The cross ratio function has been introduced by [1] and a smooth nonparametric estimator has recently been studied in [6]. As pointed out in Section 1, the CRF can be expressed in terms of the derivatives of the joint survival function. Consequently, an exact expression for $\theta(t_1, t_2)$ in (2) is given by

$$\theta(t_1, t_2) = \frac{c\left[S_1(t_1), S_2(t_2)\right] C\left[S_1(t_1), S_2(t_2)\right]}{C^{(1)}\left[S_1(t_1), S_2(t_2)\right] C^{(2)}\left[S_1(t_1), S_2(t_2)\right]}, \tag{3}$$

where $C^{(1)}(u,v) = \partial/\partial u\ C(u,v)$, $C^{(2)}(u,v) = \partial/\partial v\ C(u,v)$ and $c(u,v) = C^{(1,2)}(u,v)$ is the density function of the copula function C.

We study a simple nonparametric estimator for $\theta(t_1, t_2)$ which is obtained by replacing all quantities in (3) by empirical estimators based on a sample (T_{11}, T_{21}), …, (T_{1n}, T_{2n}) from the random vector (T_1, T_2), i.e.,

$$\widehat{\theta}(t_1, t_2) = \frac{c_{m,n}\left[S_{1n}(t_1), S_{2n}(t_2)\right] C_{m,n}\left[S_{1n}(t_1), S_{2n}(t_2)\right]}{C^{(1)}_{m,n}\left[S_{1n}(t_1), S_{2n}(t_2)\right] C^{(2)}_{m,n}\left[S_{1n}(t_1), S_{2n}(t_2)\right]}. \tag{4}$$

In the aforementioned estimator, $S_{1n}(\cdot)$ and $S_{2n}(\cdot)$ are empirical survival functions of T_1 and T_2, and $C_{m,n}$, $C^{(1)}_{m,n}$, $C^{(2)}_{m,n}$ and $c_{m,n}$ are Bernstein estimators of order m for C, $C^{(1)}$, $C^{(2)}$ and c, respectively. These estimators have been studied before in [7–9]. More specifically, we have

$$\begin{aligned}
C_{m,n}(u,v) &= \sum_{k=0}^{m}\sum_{l=0}^{m} C_n\left(\frac{k}{m}, \frac{l}{m}\right) P_{m,k}(u) P_{m,l}(v),\\
C^{(1)}_{m,n}(u,v) &= m \sum_{k=0}^{m-1}\sum_{l=0}^{m} \left[C_n\left(\frac{k+1}{m}, \frac{l}{m}\right) - C_n\left(\frac{k}{m}, \frac{l}{m}\right)\right] P_{m-1,k}(u) P_{m,l}(v),\\
C^{(2)}_{m,n}(u,v) &= m \sum_{k=0}^{m}\sum_{l=0}^{m-1} \left[C_n\left(\frac{k}{m}, \frac{l+1}{m}\right) - C_n\left(\frac{k}{m}, \frac{l}{m}\right)\right] P_{m,k}(u) P_{m-1,l}(v),\\
c_{m,n}(u,v) &= m^2 \sum_{k=0}^{m-1}\sum_{l=0}^{m-1} P_{m-1,k}(u) P_{m-1,l}(v)\ \times \qquad (5)\\
&\left[C_n\left(\frac{k+1}{m}, \frac{l+1}{m}\right) - C_n\left(\frac{k}{m}, \frac{l+1}{m}\right) - C_n\left(\frac{k+1}{m}, \frac{l}{m}\right) + C_n\left(\frac{k}{m}, \frac{l}{m}\right)\right],
\end{aligned}$$

where $C_n(u,v) = S_n\left[S_{1n}^{-1}(u), S_{2n}^{-1}(v)\right]$ represents the empirical (survival) copula function and, for $k = 0, \ldots, m$,

$$P_{m,k}(u) = \begin{pmatrix} m \\ k \end{pmatrix} u^k (1-u)^{m-k} \qquad (0 \le u \le 1),$$

are the Bernstein polynomials of order m.

The natural number m is called the order and in asymptotics we will assume that $m \to \infty$ as $n \to \infty$. See [7–9] for more details with regard to computational formulas for these expressions.

3. Asymptotic normality of the estimator

We have the following asymptotic normality result:

Theorem 3.1. *Assume*

(C1) C has bounded third order partial derivatives on $(0,1)^2$;
(C2) The copula density $c(u,v) = C^{(1,2)}(u,v)$ is Lipschitz continuous and $c \geq m_0 > 0$;
(C3) $m = Kn^{\alpha}$ with $\frac{2}{5} < \alpha < \frac{1}{2}$ and $K > 0$.

Then for all (t_1, t_2) such that $0 < S_1(t_1), S_2(t_2) < 1$, we have, as $n \to \infty$,

$$\left(\frac{n}{m}\right)^{1/2} \left(\widehat{\theta}(t_1,t_2) - \theta(t_1,t_2)\right) \xrightarrow{d} N\left(0; \frac{\theta(t_1,t_2)^2}{4\pi} \frac{1}{\sqrt{S_1(t_1)\left[1-S_1(t_1)\right]S_2(t_2)\left[1-S_2(t_2)\right]}} \frac{1}{c\left[S_1(t_1), S_2(t_2)\right]}\right).$$

Proof. We introduce shorthand notations A, C, D_1, D_2 for $c\left[S_1(t_1), S_2(t_2)\right]$, $C\left[S_1(t_1), S_2(t_2)\right]$, $C^{(1)}\left[S_1(t_1), S_2(t_2)\right]$, and $C^{(2)}\left[S_1(t_1), S_2(t_2)\right]$. Furthermore, $\widehat{A}$, $\widehat{C}$, $\widehat{D}_1$, $\widehat{D}_2$ denote $c_{m,n}\left[S_{1n}(t_1), S_{2n}(t_2)\right]$, $C_{m,n}\left[S_{1n}(t_1), S_{2n}(t_2)\right]$, $C^{(1)}_{m,n}\left[S_{1n}(t_1), S_{2n}(t_2)\right]$ and $C^{(2)}_{m,n}\left[S_{1n}(t_1), S_{2n}(t_2)\right]$, respectively.

Then $\widehat{\theta} - \theta$ can be written as follows:

$$\widehat{\theta} - \theta = \frac{\widehat{A}\widehat{C}}{\widehat{D}_1\widehat{D}_2} - \frac{AC}{D_1D_2}$$

and its asymptotic distribution will be derived from linearisation of this expression into a linear combination of $\widehat{A} - A$, $\widehat{C} - C$, $\widehat{D}_1 - D_1$ and $\widehat{D}_2 - D_2$.

Applying Lemma 4.1 in [10] gives that under condition (C1)

$$\widehat{C} - C = O_p(n^{-1/2}) + O(m^{-1}).$$

Multiplication with the scaling factor $(n/m)^{1/2}$ gives

$$\left(\frac{n}{m}\right)^{1/2} \left(\widehat{C} - C\right) = o_p(1), \tag{6}$$

if condition (C3) holds with $\alpha > 1/3$.

Applying Lemma 4.2 in [10], we have that for $j = 1, 2$:

$$\widehat{D}_j - D_j = O_p\left(\left(\frac{n}{m^{1/2}}\right)^{-1/2}\right).$$

Hence, for $j = 1, 2$,

$$\left(\frac{n}{m}\right)^{1/2} \left(\widehat{D}_j - D_j\right) = o_p(1). \tag{7}$$

The conditions needed to establish this result are (C1) and $m = Kn^{\alpha}$ with $\frac{2}{5} < \alpha < \frac{3}{5}$ (which is ensured under condition (C3)).

From equations (6) and (7) it follows that the contributions of $\widehat{C} - C$, $\widehat{D}_1 - D_1$ and $\widehat{D}_2 - D_2$ are negligible and that $\widehat{A} - A$ will determine the asymptotic behaviour of

$\widehat{\theta} - \theta$:

$$\widehat{\theta} - \theta \sim \frac{C}{D_1 D_2}\left(\widehat{A} - A\right) = \frac{\theta}{A}\left(\widehat{A} - A\right)$$

It remains to establish the asymptotic normality of $\widehat{A} - A$. An application of the Lemma in Appendix A gives that

$$\begin{aligned}\widehat{A} - A &= c_{m,n}\left[S_1(t_1), S_2(t_2)\right] - c\left[S_1(t_1), S_2(t_2)\right] \\ &\quad + O_p\left(n^{(3\alpha/2)-1}\left(\log n\right)^{1/2}\left(\log\log n\right)^{3/4} + m^{-1} + n^{-1/2}\right).\end{aligned}$$

After multiplication with $(n/m)^{1/2}$, the O_p-term tends to zero in probability if $\alpha < 1/2$. Therefore, $\widehat{A} - A$ has the same asymptotic distribution as $c_{m,n}\left[S_1(t_1), S_2(t_2)\right] - c\left[S_1(t_1), S_2(t_2)\right]$. For the latter, we have the asymptotic normality result in [8].

From this the theorem follows. □

Remark 1. The difference in terms of asymptotic variance between our novel estimator and the smooth estimator studied in [6] is best visible in the denominator of the expression of the asymptotic variance. Kernel smoothing in the estimator in [6] implies the presence of marginal density functions, say $f_1(t_1)$ and $f_2(t_2)$, related to T_1 and T_2, respectively. This differs from the denominator in the asymptotic variance in Theorem 3.1, in which, for example, $f_1(t_1)$ is now replaced by the term $\sqrt{(S_1(t_1)[1 - S_1(t_1)]}$. A comparison between these two quantities is possible thanks to a result of Parzen [11]. From that result, it follows that $\sqrt{(S_1(t_1)[1 - S_1(t_1)]}$ is asymptotically larger than $f_1(t_1)$ (i.e., as $t_1 \to \infty$) for all random variables T_1 with medium tails (e.g., exponential, Weibull or normal distributed random variables) and long tails (e.g., Cauchy or Pareto distributions) (see also Remark 5 in [9]). Similarly, for $f_2(t_2)$ and $\sqrt{(S_2(t_2)[1 - S_2(t_2)]}$ the same holds. Consequently, this leads to a smaller asymptotic variance for the novel estimator as compared to the asymptotic variance of the estimator proposed by [6].

4. Simulation study

Based on simulations we show the finite sample performance of our estimator $\widehat{\theta}(t_1, t_2)$ in equation (4). First, we describe the simulation procedure after which we summarize the simulation results.

4.1. Simulation procedure

We generate n pairs of event times (t_{1j}, t_{2j}), $j = 1, \ldots, n$ using the *Copula.surv*-package in R. More specifically, random samples (u_{1j}, u_{2j}) are drawn from three different copula functions with various tail dependencies (independence, Clayton and Gumbel copula function) after which dependent exponentially distributed event times are obtained with rate parameters equal to $\lambda_1 = 0.03$ and $\lambda_2 = 0.05$ for T_1 and T_2, respectively, as follows:

$$t_{ij} = -\frac{\ln(u_{ij})}{\lambda_i}.$$

The Clayton copula captures lower tail dependence, while the Gumbel copula captures upper tail dependence. In our simulation study, we generate simulation sets of sample size $n = 500$. However, additional simulation results for varying sample sizes and copula functions are provided in Appendix A of the Supplementary Material. We consider $m = Kn^{\alpha}$, with $\alpha = 9/20$ (i.e., the average of the theoretical bounds on α) and $K = 2$. The impact of considering different K-values is illustrated in Appendix A of the Supplementary Material.

4.2. Independence copula

First, we depict the simulation results under the assumption of independence of T_1 and T_2. More specifically, we generate 100 simulation sets of size $n = 500$ under independence and estimate the cross ratio function with the true cross ratio being constant and equal to one. In Figure 1, we graphically show a heatplot of the difference between the estimated cross ratio values $\widehat{\theta}_m(t_1, t_2)$ averaged over the simulation runs and the true values $\theta(t_1, t_2)$ (left upper panel), and the estimated cross ratio function $\widehat{\theta}_m(t_1, t_2)$ (black solid lines in the other panels) as a function of one time component by fixing the other ($t_1 = F_1^{-1}(0.5)$ in the right upper panel, or $t_2 = F_2^{-1}(0.5)$ in the left lower panel, respectively). Furthermore, the Bernstein order is taken equal to $m = 2n^{9/20}$ with sample size $n = 500$ in each of the simulation runs. Pointwise 95% simulation-based confidence bands (gray shaded areas) and true cross ratio values (red dashed lines) are included as well. In the right lower panel, we plot the cross ratio function $\widehat{\theta}_m[F_1^{-1}(u), F_2^{-1}(u)]$ against $u \in (0, 1)$. Overall, the estimator is performing well. Clearly the simulation-based variability is larger for small t_1 and/or t_2-values as depicted in Figure 1. This is as expected based on the expression for the asymptotic variance (see Theorem 3.1) in which the denominator will become small in the aforementioned situation.

4.3. Clayton copula

We now consider the Clayton copula function with parameter $\theta = 0.5$. Consequently, the true underlying cross ratio function takes constant value $1 + \theta \; = \; 1.5$. In Figure 2, we graphically show a heatplot of the difference between the estimated cross ratio values $\widehat{\theta}_m(t_1, t_2)$ averaged over the 100 simulation runs and the true values $\theta(t_1, t_2)$ (left upper panel), and the estimated cross ratio function $\widehat{\theta}_m(t_1, t_2)$ (black solid lines in the other panels) as a function of one time component by fixing the other ($t_1 = F_1^{-1}(0.5)$ in the right upper panel, or $t_2 = F_2^{-1}(0.5)$ in the left lower panel, respectively). Furthermore, the Bernstein order is taken equal to $m = 2n^{9/20}$ with sample size $n = 500$ in each of the simulation runs. Pointwise 95% simulation-based confidence bands (gray shaded areas) and true cross ratio values (red dashed lines) are included as well. In the right lower panel, we plot the cross ratio function $\widehat{\theta}_m[F_1^{-1}(u), F_2^{-1}(u)]$ against $u \in (0, 1)$. In general, the estimator performs well in terms of estimating the true underlying cross ratio surface, except for $(t_1, t_2)-$values corresponding to higher quantiles. This can be explained by the fact that a Clayton copula function implies lower tail dependence, hence, estimation is more difficult for higher quantiles.

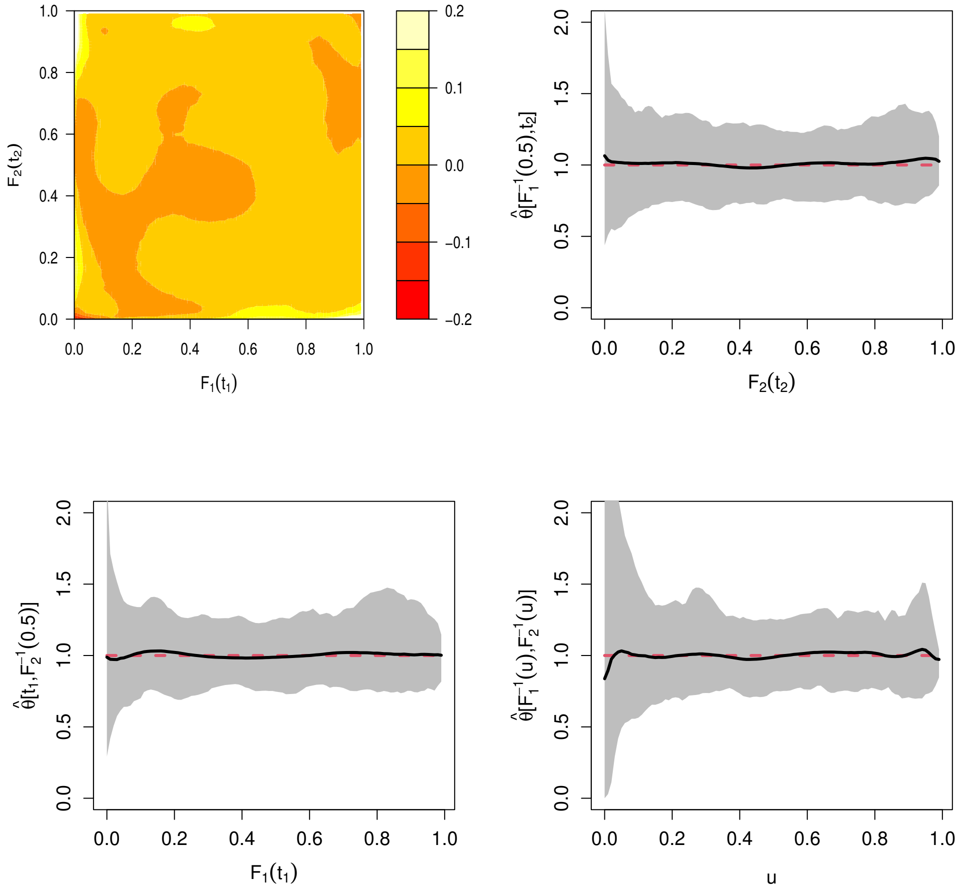

Figure 1. Independence copula cross ratio function estimation with $m = 2n^{\alpha}$, $\alpha = 9/20$ and $n = 500$: heatplot representing the difference between the estimated cross ratio function $\widehat{\theta}_m(t_1, t_2)$ averaged over 100 simulation runs and the true cross ratio function $\theta(t_1, t_2)$ (left upper panel) and intersections of the estimated cross ratio surface given $t_1 = F_1^{-1}(0.5)$ (right upper panel; black solid line), $t_2 = F_2^{-1}(0.5)$ (left lower panel) and $F_1^{-1}(u) = F_2^{-1}(u)$ (right lower panel) with pointwise 95% simulation-based confidence bands (gray region). True cross ratio curves are graphically depicted in red dashed lines.

4.4. Gumbel copula

Finally a Gumbel copula function is considered with association parameter $\theta = 1.25$ and expression for the cross ratio function equal to

$$\theta(t_1, t_2) = 1 + (\theta - 1)\left(\{-\ln[S_1(t_1)]\}^{\theta} + \{-\ln[S_2(t_2)]\}^{\theta}\right)^{-1/\theta}.$$

In Figure 3, we graphically depict the difference between the average estimated cross ratio function $\widehat{\vartheta}_m(t_1, t_2)$ and the true cross ratio function (heatplot in left upper panel). Intersections of the averaged estimated cross ratio function (black solid lines) are shown together with pointwise 95% simulation-based confidence bands for $m = 2n^{9/20}$ and $n = 500$. Although on average the nonparametric estimator for the cross ratio is in close to the true CRF, $\theta(t_1, t_2)$ is slightly underestimated for small values of (t_1, t_2) in the lower left corner of the surface (see white regions on the heatplot). Here, the Gumbel

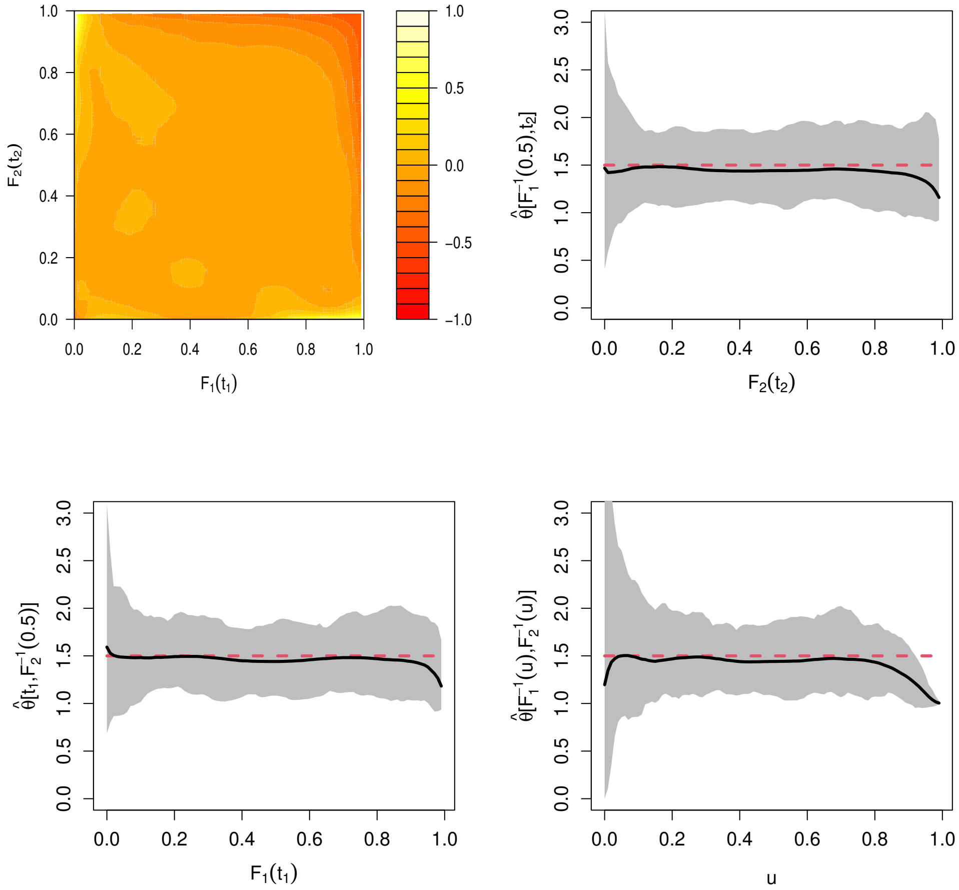

Figure 2. Clayton copula cross ratio function estimation with $m = 2n^{\alpha}$, $\alpha = 9/20$ and $n = 500$: heatplot representing the difference between the estimated cross ratio function $\widehat{\theta}_m(t_1, t_2)$ averaged over 100 simulation runs and the true cross ratio function $\theta(t_1, t_2)$ (left upper panel) and intersections of the estimated cross ratio surface given $t_1 = F_1^{-1}(0.5)$ (right upper panel; black solid line), $t_2 = F_2^{-1}(0.5)$ (left lower panel) and $F_1^{-1}(u) = F_2^{-1}(u)$ (right lower panel) with pointwise 95% simulation-based confidence bands (gray region). True cross ratio curves are graphically depicted in red dashed lines.

copula function gives rise to upper tail dependence thereby implying difficulties when estimating the cross ratio function for lower quantiles.

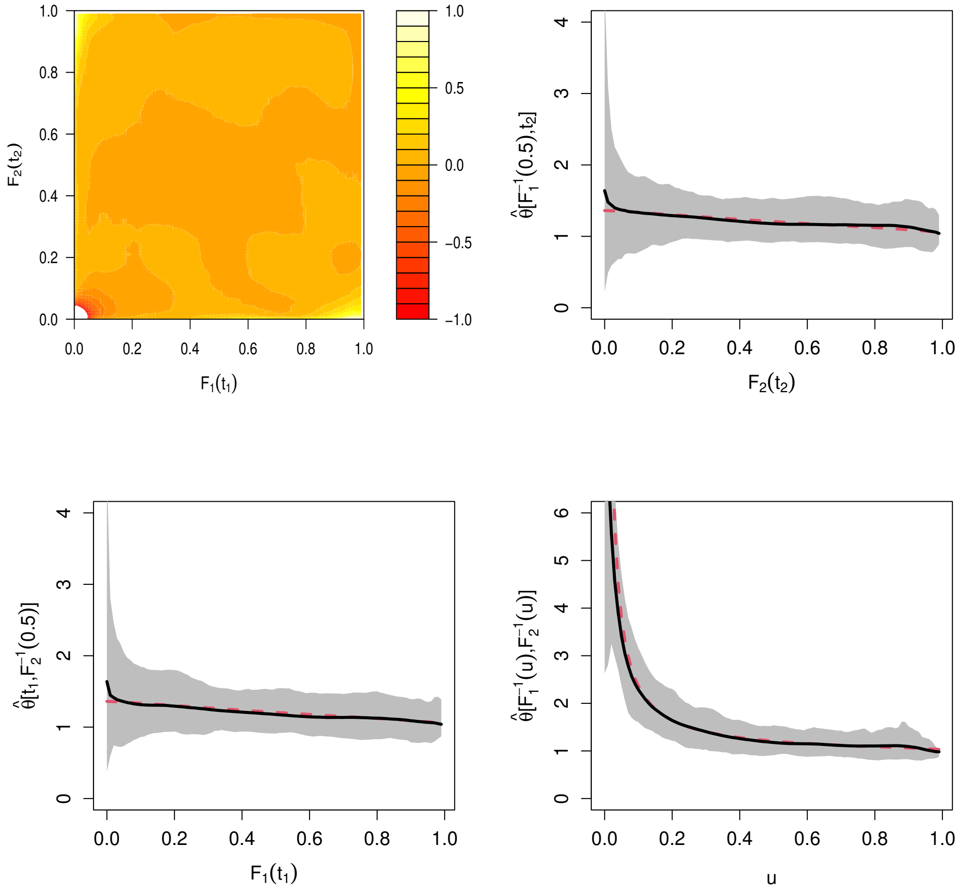

Figure 3. Gumbel copula cross ratio function estimation with $m = 2n^{\alpha}$, $\alpha = 9/20$ and $n = 500$: heatplot representing the difference between the estimated cross ratio function $\widehat{\theta}_m(t_1, t_2)$ averaged over 100 simulation runs and the true cross ratio function $\theta(t_1, t_2)$ (left upper panel) and intersections of the estimated cross ratio surface given $t_1 = F_1^{-1}(0.5)$ (right upper panel; black solid line), $t_2 = F_2^{-1}(0.5)$ (left lower panel) and $F_1^{-1}(u) = F_2^{-1}(u)$ (right lower panel) with pointwise 95% simulation-based confidence bands (gray region). True cross ratio curves are graphically depicted in red dashed lines.

5. Data application

The use of the estimator $\widehat{\theta}(t_1, t_2)$ for the cross ratio function is demonstrated based on hospital data collected amidst the first wave of the Belgian COVID-19 epidemic. The hospitalization data used in this paper consists of patient information with regard to patients admitted to Ziekenhuis Oost Limburg (ZOL), Genk, Limburg, Belgium, after severe infection with the novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). More specifically, the data were collected during the first COVID-19 wave in Belgium with patients admitted to ZOL between March 16, 2020 and May 30, 2020. This study was approved by the Medical Ethical Comittee of ZOL.

In total, 300 patients were hospitalized of which 47 died in the hospital. In the further analysis, we confine attention to the $n = 253$ patients that recovered and have been discharged from the hospital.

The presence of SARS-CoV-2 is determined using a semi-quantitative RT-PCR test (Allplex$^{\text{TM}}$ 2019-nCoV Assay, Seegene, Seoul, Korea), a molecular technique to detect a selection of genes (E-gene, RdRP, N-gene) related to the virus, following an oro-nasopharyngeal swab. The length of stay in the hospital (before discharge or death) for SARS-CoV-2 infected patients is studied in relation to their age. In Figure 4, we graphically depict the length of stay in the hospital in relation to the age of individuals.

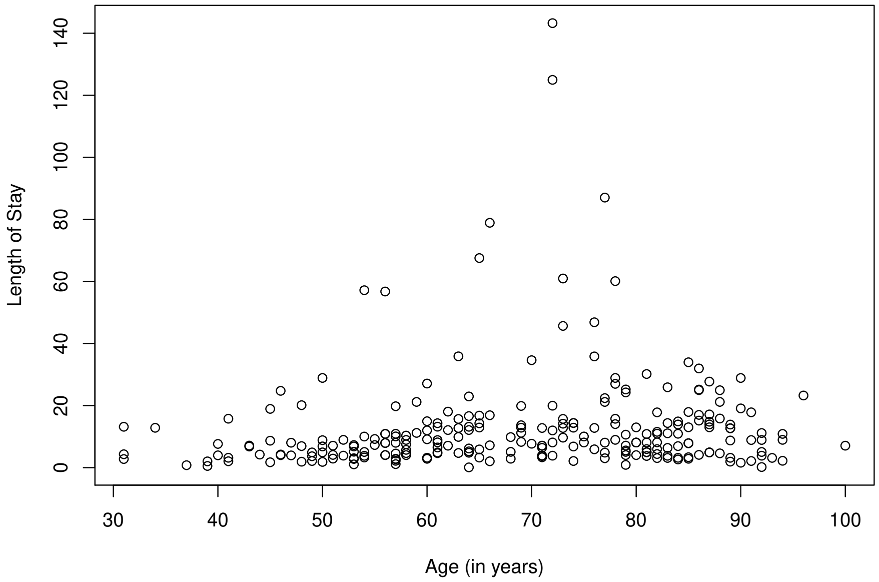

Figure 4. Scatterplot of the observed length of stay in relation to the age of patients hospitalized in Ziekenhuis Oost Limburg (ZOL) during the first COVID-19 wave.

The global association between age at admission and hospital length of stay, expressed in terms of Kendall's tau ($\tau = 0.149$, 95% asymptotic confidence interval: $[0.065, 0.232]$), is estimated to be relatively small though positive. This implies that older individuals tend to have larger recovery times before leaving the hospital.

In order to measure the local strength of association between the age of the patient and the length of stay in the hospital following COVID-19 infection, we estimate the cross-ratio function $\widehat{\theta}(t_1, t_2)$, where T_1 represents the length of stay in the hospital and T_2 denotes the age of a patient at hospital admission. In Figure 5, we show the estimated cross-ratio surface (with $m = 2n^{\alpha}$, where $\alpha = 0.45$) for the dependence between the age of hospitalized COVID-19 positive patients and their length of stay in the hospital. Clearly, the strength of association is highest in the left lower corner of the surface plot, thereby indicating that the association between length of stay and age is strongest among younger patients.

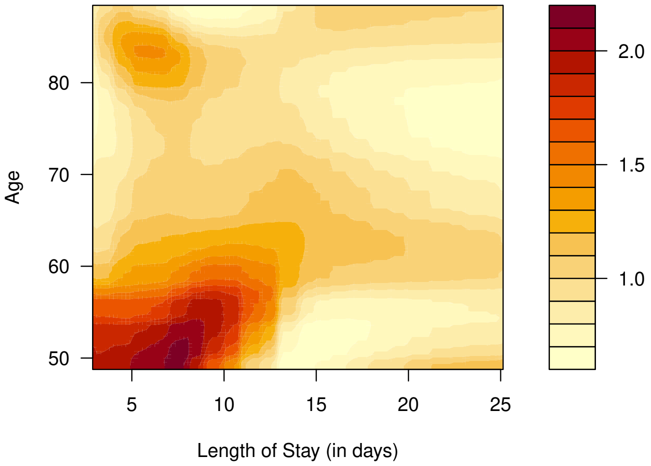

Figure 5. Hospital data application: Cross-ratio function surface showing the relation between age at admission and hospital length of stay for recovered COVID-19 patients.

In order to estimate the variability associated with the estimation of the cross ratio function, a bootstrap approach is considered in which the data is resampled nonparametrically. In total, $B = 500$ bootstrap samples have been generated and the cross ratio estimate, denoted by $\widehat{\theta}^{(b)}(t_1, t_2)$, was obtained based on boostrap sample $b = 1, \ldots, B$. In Figure 6, we present the estimated cross ratio curves for patients of age $t_2 = 50$ years (upper left panel), $t_2 = 60$ years (upper right panel), $t_2 = 70$ years (lower left panel), or $t_2 = 80$ years (lower right panel). Pointwise 95% bootstrap-percentile confidence bands are shown as gray shaded areas. A strong positive local association is observed for relatively young patients thereby implying a shorter recovery time as compared to older patients (cfr. hazard interpretation of cross ratio function). For older patients, the cross ratio values are not significantly different from one across all hospital lengths of stay, thereby leading to the conclusion that the local association between age and length of stay vanishes for older patients. Essentially, the discharge rate stabilizes for older patients (aged > 60 years) as compared to the oldest patients in the sample.

6. Discussion

In this paper we propose a Bernstein-based estimator for the CRF which is an alternative to the smooth estimator studied in Abrams et al. [6]. The choice for Bernstein-based estimators is made therein because of their well known good bias and variance

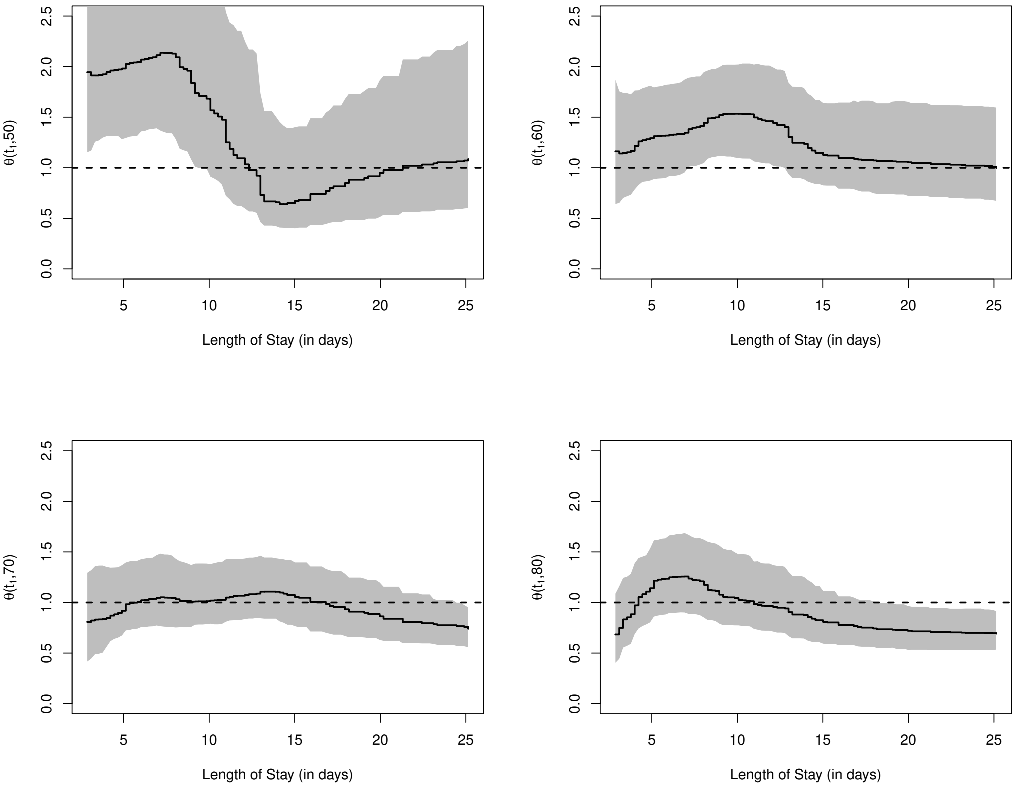

Figure 6. Hospital data application: Cross ratio curves (black solid lines) for the age of individuals equal to $t_2 = 50$ years (upper left panel), $t_2 = 60$ years (upper right panel), $t_2 = 70$ years (lower left panel), or $t_2 = 80$ years (lower right panel) together with pointwise 95% bootstrap-percentile confidence bands (gray shaded area).

properties, in particular the absence of boundary effects. More specifically, this work is based on the seminal paper by Leblanc [15] on Bernstein estimation for a cumulative distribution function on $[0, 1]$ and on more recent work by Janssen et al. [7] with respect to the use of Bernstein polynomials in copula estimation. In Ouimet [16] there is, next to an excellent overview of the Bernstein literature, a generalization to the d-dimensional simplex. For distribution functions on $[0, \infty)$, Bernstein estimation with Poisson weights instead of binomial weights has been considered by Chaubey et al. [17,18] for smooth estimation of univariate and multivariate survival and density functions.

In Abrams et al. [6], the conditional hazard rate functions in the definition (5) of the CRF are estimated by first applying Bernstein methods to the cumulative hazard rate functions, followed by a kernel smoothing approach. Our new proposal uses Bernstein estimators for the four components of the CRF in (4), and avoiding a further smoothing step. A direct comparison between our estimator and the one proposed by Abrams et al. [6] is complicated by different restrictions in terms of the Bernstein order and bandwidth related to the latter one.

Our new estimator $\widehat{\theta}(t_1, t_2)$ in (4) is not smooth due to the presence of the empirical survival functions $S_{1n}(t_1)$ and $S_{2n}(t_2)$ for the marginal survival functions. They can be replaced by kernel survival function estimators $\widetilde{S}_{1n}(t_1)$ and $\widetilde{S}_{2n}(t_2)$ defined, for

$j = 1, 2$, as

$$\widetilde{S}_{jn}(t_j) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{K}\left(\frac{T_{ji} - t_j}{h_n}\right),$$

where $\mathbb{K}$ is a known distribution function having a density function K that is continuous and symmetric about zero and $h_n > 0$ is a bandwidth sequence with $h_n \to 0$ as $n \to \infty$.

It can be shown in a similar way that the same asymptotic normality result is valid under the extra assumptions that T_1 and T_2 have densities f_1 and f_2 with f_1' and f_2' bounded and that $h_n = n^{-\beta}$ with $\beta \geq 1/4$.

Further smoothing of the different components in the cross ratio function and the corresponding asymptotic normality result is available though not shown in this paper.

The proposed estimators of the cross ratio function are derived under the assumptions of no censoring. Extensions to randomly right-censored data are currently under investigation. This requires new results for copulas and derivatives under different types of censoring. Relevant references are Khardani [19] for distributions on $[0, 1]$ and Geerdens et al. [20] for copula estimation under different types of bivariate censoring.

A challenging problem is the estimation of the optimal Bernstein polynomial order m which is closely related to the finding the optimal kernel bandwidth in kernel-based estimation methods. In the literature, plug-in and cross-validation approaches have been discussed for determining the optimal bandwidths in kernel-based estimation of the density or distribution function. These approaches typically consider the (asymptotic) mean integrated squared error as a criterion to study the trade-off between bias and variance. Despite the use of least-squares leave-one-out cross-validation in univariate and multivariate kernel density estimation, it has been questioned in the context of distribution estimation [13,14]. A detailed study of different bandwidth selection methods is therefore required in the context of the estimation of the cross ratio function. An interesting open research question is to explore minimax properties of the proposed CRF estimator. A recent reference from the extensive literature on minimax estimation is particularly relevant here, i.e., Bertin et al. [21].

Disclosure statement

No potential conflict of interest was reported by the author(s).

Acknowledgements

The authors express their gratitude to access and use the hospital data collected in Ziekenhuis Oost Limburg to illustrate the novel methodology presented in this paper.

Funding

SA and OS gratefully acknowledge support of the Research Foundation - Flanders (FWO) (grant nr.G011022N).

References

[1] Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65, 141–151.

[2] Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society Series B – Statistical Methodology*, 44, 414–422.

[3] Oakes, D. (1986). Semi-parametric inference in a model for association in bivariate survival data. *Biometrika*, 73, 353–361.

[4] Oakes, D. (1989). Bivariate survival data induced by frailties. *Journal of the American Statistical Association*, 84, 487–493.

[5] Hu, T., Nan, B., Lin, X., Robins, J. M. (2011). Time-dependent cross ratio estimation for bivariate failure times. *Biometrika*, 98, 341–354.

[6] Abrams, S., Janssen, P., Swanepoel, J., Veraverbeke, N. (2020). Nonparametric estimation of the cross ratio function. *Annals of the Institute of Statistical Mathematics*, 72(3), 771–801.

[7] Janssen, P., Swanepoel, J., Veraverbeke, N. (2012). Large sample behavior of the Bernstein copula estimator. *Journal of Statistical Planning and Inference*, 142, 1189–1197.

[8] Janssen, P., Swanepoel, J., Veraverbeke, N. (2014). A note on the asymptotic behavior of the Bernstein estimator of the copula density. *Journal of Multivariate Analysis*, 124, 480–487.

[9] Janssen, P., Swanepoel, J., Veraverbeke, N. (2016). Bernstein estimation for a copula derivative with application to conditional distribution and regression functionals. *Test*, 25, 351–374.

[10] Abrams, S., Janssen, P., Swanepoel, J., Veraverbeke, N. (2022). Nonparametric estimation of risk ratios for bivariate data. *Statistics*, 34(4), 940–963.

[11] Parzen, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association*, 74, 105–121.

[12] Stute, W. (1984). The oscillation behavior of empirical processes: The multivariate case *The Annals of Probability*, 12(2), 361–379.

[13] Altman, N., Léger, C. (1995). Bandwidth selection for kernel distribution function estimation *Journal of Statistical Planning and Inference*, 46, 195–214.

[14] Bowman, A., Hall, P., Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, 85, 799–808.

[15] Leblanc A. (2012). On estimating distribution functions using Bernstein polynomials. *Annals of the Institute of Statistical Mathematics*, 64, 919–943.

[16] Ouimet, F. (2021). Asymptotic properties of Bernstein estimators on the simplex. *Journal of Multivariate Analysis*, 185, 104784.

[17] Chaubey Y. P., Sen P. K. (1996). On smooth estimation of survival and density functions. *Statistics and Decisions*, 14, 1–22.

[18] Chaubey Y. P., Sen P. K. (1996). Smooth estimation of multivariate survival and density functions. *Journal of Statistical Planning and Inference*, 103, 361–376.

[19] Khardani S. (2023). A Bernstein polynomial approach to the estimation of a distribution function and quantiles under censorship model. *Communications in Statistics - Theory and Methods*. doi: 10.1080/03610926.2023.2228948.

[20] Geerdens C., Janssen P., Veraverbeke N.(2016). Large sample properties of nonparametric copula estimators under bivariate censoring. *Statistics*, 50, 1036–1055.

[21] Bertin K., Genest C, Klutchendorf F., Quimet F. (2023). Minimax properties of Dirichlet kernel density estimators. *Journal of Multivariate Analysis*, 195, 105158.

Appendix A.

Lemma. *If the second order partial derivatives $C^{(1,1)}$, $C^{(2,2)}$ and $C^{(1,2)} = c$ exist and are continuous on $[0,1]^2$, if c is Lipschitz continuous and if $m = Kn^{\alpha}$, $\frac{1}{3} < \alpha < \frac{1}{2}$, $K > 0$, then*

$$\begin{aligned} & c_{m,n}\left[S_{1n}(t_1), S_{2n}(t_2)\right] - c\left[S_1(t_1), S_2(t_2)\right] \\ &= c_{m,n}\left[S_1(t_1), S_2(t_2)\right] - c\left[S_1(t_1), S_2(t_2)\right] \\ & \quad + O_p\left(n^{(3\alpha/2)-1}\left(\log n\right)^{1/2}\left(\log\log n\right)^{3/4} + m^{-1} + n^{-1/2}\right). \end{aligned}$$

Proof. Define

$$\begin{aligned} b_m(u,v) =& m^2 \sum_{k=0}^{m-1}\sum_{l=0}^{m-1} P_{m-1,k}(u)P_{m-1,l}(v) \times \\ & \left[C\left(\frac{k+1}{m}, \frac{l+1}{m}\right) - C\left(\frac{k}{m}, \frac{l+1}{m}\right) - C\left(\frac{k+1}{m}, \frac{l}{m}\right) + C\left(\frac{k}{m}, \frac{l}{m}\right)\right]. \end{aligned}$$

From this and the expression of $c_{m,n}(u,v)$ in (5), we have:

$$\begin{aligned} & c_{m,n}\left[S_{1n}(t_1), S_{2n}(t_2)\right] - c\left[S_1(t_1), S_2(t_2)\right] \\ &= c_{m,n}\left[S_{1n}(t_1), S_{2n}(t_2)\right] - b_m\left[S_{1n}(t_1), S_{2n}(t_2)\right] \\ & \qquad - \left\{c_{m,n}\left[S_1(t_1), S_2(t_2)\right] - b_m\left[S_1(t_1), S_2(t_2)\right]\right\} \\ & \quad + c_{m,n}\left[S_1(t_1), S_2(t_2)\right] - c\left[S_1(t_1), S_2(t_2)\right] \\ & \quad + b_m\left[S_{1n}(t_1), S_{2n}(t_2)\right] - b_m\left[S_1(t_1), S_2(t_2)\right] \\ &=: (I) + (II) + (III). \end{aligned}$$

From the proof of the Theorem in [8], we obtain

$$(III) = c\left[S_{1n}(t_1), S_{2n}(t_2)\right] - c\left[S_1(t_1), S_2(t_2)\right] + O(m^{-1}),$$

and since c is Lipschitz continuous

$$(III) = O_p(n^{-1/2}) + O(m^{-1}).$$

We now deal with term (I):

$$
\begin{aligned}
(I) = m^2 \sum_{k=0}^{m-1} \sum_{l=0}^{m-1} \Bigg\{ & \left[C_n \left(\frac{k+1}{m}, \frac{l+1}{m} \right) - C \left(\frac{k+1}{m}, \frac{l+1}{m} \right) \right] - \\
& \left[C_n \left(\frac{k}{m}, \frac{l+1}{m} \right) - C \left(\frac{k}{m}, \frac{l+1}{m} \right) \right] - \\
& \left[C_n \left(\frac{k+1}{m}, \frac{l}{m} \right) - C \left(\frac{k+1}{m}, \frac{l}{m} \right) \right] + \\
& \left[C_n \left(\frac{k}{m}, \frac{l}{m} \right) - C \left(\frac{k}{m}, \frac{l}{m} \right) \right] \Bigg\} \times \\
& \Big\{ P_{m-1,k} \left[S_{1n}(t_1) \right] P_{m-1,l} \left[S_{2n}(t_2) \right] - P_{m-1,k} \left[S_1(t_1) \right] P_{m-1,l} \left[S_2(t_2) \right] \Big\} \\
= \frac{m^2}{n^{1/2}} \sum_{k=0}^{m-1} \sum_{l=0}^{m-1} & \left[\alpha_n^C \left(\frac{k+1}{m}, \frac{l+1}{m} \right) - \alpha_n^C \left(\frac{k}{m}, \frac{l+1}{m} \right) - \right. \\
& \left. \alpha_n^C \left(\frac{k+1}{m}, \frac{l}{m} \right) + \alpha_n^C \left(\frac{k}{m}, \frac{l}{m} \right) \right] \times \\
& \big\{ \left[S_{1n}(t_1) - S_1(t_1) \right] P'_{m-1,k} \left(\theta_1 \right) P_{m-1,l} \left(\theta_2 \right) + \\
& \quad \left[S_{2n}(t_2) - S_2(t_2) \right] P_{m-1,k} \left(\theta_1 \right) P'_{m-1,l} \left(\theta_2 \right) \big\},
\end{aligned}
$$

where (θ_1, θ_2) lies between $(S_{1n}(t_1), S_{2n}(t_2))$ and $(S_1(t_1), S_2(t_2))$, and where

$$
\alpha_n^C(u,v) = n^{1/2} \left[C_n(u,v) - C(u,v) \right]
$$

is the empirical copula process.

For a rectangle R in $[0,1]^2$ we denote

$$
\alpha_n^C(R) = n^{1/2} \left[\mu_n^C(R) - \mu^C(R) \right],
$$

where μ_n^C and μ^C are the measures corresponding to the distribution functions C_n and C.

Also denote, for $0 \leq k, l \leq m-1$:

$$
R_{k,l} = \left] \frac{k}{m}, \frac{k+1}{m} \right] \times \left] \frac{l}{m}, \frac{l+1}{m} \right].
$$

Then,

$$
\begin{aligned}
(I) = \frac{m^2}{n^{1/2}} \sum_{k=0}^{m-1} \sum_{l=0}^{m-1} & \alpha_n^C \left(R_{k,l} \right) \times \\
& \big\{ \left[S_{1n}(t_1) - S_1(t_1) \right] P'_{m-1,k} \left(\theta_1 \right) P_{m-1,l} \left(\theta_2 \right) + \\
& \quad \left[S_{2n}(t_2) - S_2(t_2) \right] P_{m-1,k} \left(\theta_1 \right) P'_{m-1,l} \left(\theta_2 \right) \big\}. \qquad \text{(A1)}
\end{aligned}
$$

We now use an almost sure representation of Stute [12], i.e., if the second order partial derivatives of C exist and are continuous on $[0,1]^2$, then uniformly

$$\begin{aligned} C_n(u,v) - C(u,v) = & \frac{1}{n}\sum_{i=1}^{n}\left[I(U_i \le u, V_i \le v) - C(u,v)\right] - \\ & C^{(1)}(u,v)\left[\frac{1}{n}\sum_{i=1}^{n} I(U_i \le u) - u\right] - \\ & C^{(2)}(u,v)\left[\frac{1}{n}\sum_{i=1}^{n} I(V_i \le v) - v\right] + \\ & O\left(n^{-3/4}\left(\log n\right)^{1/2}\left(\log\log n\right)^{1/4}\right) \quad \text{a.s.} \end{aligned}$$

Here the vectors $(U_1, V_1), \ldots, (U_n, V_n)$ are independent with common joint survival function C.

Introduce

$$\begin{aligned} G_n(u,v) &= \frac{1}{n}\sum_{i=1}^{n} I(U_i \le u, V_i \le v) \\ G(u,v) &= C(u,v) \end{aligned}$$

Also, the two-dimensional empirical process

$$\alpha_n^G(u,v) = n^{1/2}\left[G_n(u,v) - G(u,v)\right]$$

and the one-dimensional empirical processes

$$\begin{aligned} \alpha_{1n}^G(u) &= n^{1/2}\left[G_n(u,1) - G(u,1)\right] \\ \alpha_{2n}^G(v) &= n^{1/2}\left[G_n(1,v) - G(1,v)\right]. \end{aligned}$$

From the result of Stute [12] we then have

$$\begin{aligned} \alpha_n^C(R_{k,l}) = & \alpha_n^G(R_{k,l}) - \\ & \left[C^{(1)}\left(\frac{k+1}{m}, \frac{l+1}{m}\right) - C^{(1)}\left(\frac{k+1}{m}, \frac{l}{m}\right)\right]\alpha_{1n}^G\left(\frac{k+1}{m}\right) \\ & + \left[C^{(1)}\left(\frac{k}{m}, \frac{l+1}{m}\right) - C^{(1)}\left(\frac{k}{m}, \frac{l}{m}\right)\right]\alpha_{1n}^G\left(\frac{k}{m}\right) \\ & - \left[C^{(2)}\left(\frac{k+1}{m}, \frac{l+1}{m}\right) - C^{(2)}\left(\frac{k}{m}, \frac{l+1}{m}\right)\right]\alpha_{2n}^G\left(\frac{l+1}{m}\right) \\ & + \left[C^{(2)}\left(\frac{k+1}{m}, \frac{l}{m}\right) - C^{(2)}\left(\frac{k}{m}, \frac{l}{m}\right)\right]\alpha_{2n}^G\left(\frac{l}{m}\right) \\ & + O\left(n^{-1/4}\left(\log n\right)^{1/2}\left(\log\log n\right)^{1/4}\right) \quad \text{a.s.} \end{aligned}$$

Now use that $C^{(1)}$ and $C^{(2)}$ are Lipschitz continuous and that $\sup_{0 \le u \le 1} |\alpha_{1n}^G(u)|$ and

$\sup_{0\le v\le 1}|\alpha_{2n}^G(v)|$ are $O\left((\log\log n)^{1/2}\right)$ a.s. to obtain

$$\sup_{0\le k,l\le m-1}|\alpha_n^C(R_{k,l})| \le \sup_{0\le k,l\le m-1}|\alpha_n^G(R_{k,l})| + O\left(m^{-1}(\log\log n)^{1/2}\right) + O\left(n^{-1/4}(\log n)^{1/2}(\log\log n)^{1/4}\right) \quad \text{a.s.} \tag{A2}$$

For the first term in the right hand side of this inequality, we recall a result of Stute [12] on the oscillation behaviour of the multivariate empirical process. For the empirical process α_n^G we define the oscillation modulus as

$$\omega_n(a_1,a_2) = \sup_{y_1-x_1\le a_1, y_2-x_2\le a_2}|\alpha_n^G(R_{\underset{\sim}{x},\underset{\sim}{y}})|,$$

where $R_{\underset{\sim}{x},\underset{\sim}{y}} =]x_1,y_1]\times]x_2,y_2]$.

Theorem 2 of Stute [12] says: if C has a continuous density c on $[0,1]^2$ with $c\ge m_0>0$ and if (a_n^2) is a *bandsequence*, then

$$\omega_n(a_n,a_n) = O\left(\sqrt{a_n^2\log\left(\frac{1}{a_n^2}\right)}\right) \quad \text{a.s.}$$

We have

$$\sup_{0\le k,l\le m-1}|\alpha_n^G(R_{k,l})| \le \omega\left(\frac{1}{m},\frac{1}{m}\right). \tag{A3}$$

Apply Theorem 2.1 of Stute [12] to the last expression. Here $a_n = m^{-1}$ and we can check that $(a_n^2) = (m^{-2})$ is a bandsequence:

(i) $na_n^2 = nm^{-2} = n^{1-2\alpha}$ if $m = n^\alpha$ and this tends to $+\infty$ if $\alpha < \frac{1}{2}$;

(ii) $\log\left(\frac{1}{a_n^2}\right) = 2\alpha\log n = o(n^{1-2\alpha})$;

(iii) $\log\left(\frac{1}{a_n^2}\right)/\log\log n \to \infty$.

Hence,

$$\omega_n\left(\frac{1}{m},\frac{1}{m}\right) = O\left(m^{-1}(\log n)^{1/2}\right) \quad \text{a.s.} \tag{A4}$$

From (A2), (A3) and (A4)

$$\sup_{0\le k,l\le m-1}|\alpha_n^C(R_{k,l})| = O\left(m^{-1}(\log n)^{1/2}\right) + O\left(n^{-1/4}(\log n)^{1/2}(\log\log n)^{1/4}\right) \quad \text{a.s.}$$

This, combined with (A1) gives that, a.s.,

$$(I) = O\left(\frac{m^2}{n^{1/2}}\left(m^{-1}(\log n)^{1/2} + n^{-1/4}(\log n)^{1/2}(\log\log n)^{1/4}\right)\left(n^{-1/2}(\log\log n)^{1/2}m^{1/2}\right)\right).$$

The last factor $m^{1/2}$ comes from the fact that $\sum_{k=0}^{m-1} |P'_{m-1,k}(\theta_1)|$ and $\sum_{l=0}^{m-1} |P'_{m-1,l}(\theta_2)|$ are $O(m^{1/2})$ (see Lemma 1 in [8]).

With $m = Kn^{\alpha}$ this becomes

$$\begin{aligned}(I) &= O\left(n^{(3\alpha/2)-1} (\log n)^{1/2} (\log \log n)^{1/2} + n^{(5\alpha/2)-5/4} (\log n)^{1/2} (\log \log n)^{3/4}\right) \quad \text{a.s.} \\ &= O\left(n^{(3\alpha/2)-1} (\log n)^{1/2} (\log \log n)^{3/4}\right) \quad \text{a.s.,}\end{aligned}$$

since $\dfrac{3\alpha}{2} - 1 < \dfrac{5\alpha}{2} - \dfrac{5}{4} < 0$ for $\dfrac{1}{4} < \alpha < \dfrac{1}{2}$. □