Contents lists available at ScienceDirect

C. R. Acad. Sci. Paris, Ser. I

www.sciencedirect.com

Statistics

A semiparametric test of independence in copula models for censored data

Test d'indépendance semiparamétrique dans des modèles de copule pour les données censurées

Salim Bouzebda [a], Amor Keziou [b,a]

[a] *LSTA-université Paris 6, 175, rue du Chevaleret, boîte 158, 75013 Paris, France*
[b] *Laboratoire de mathématiques (FRE 3111) CNRS, université de Reims, Reims, France*

ARTICLE INFO

Article history:
Received 15 April 2009
Accepted after revision 7 February 2010

Presented by Wendelin Werner

ABSTRACT

We propose a semiparametric test of independence in copula models for bivariate survival censored data. We give the limit laws of the estimate of the parameter and the proposed test statistic under the null hypothesis of independence.

RÉSUMÉ

Nous proposons un test d'indépendance dans des modèles de copule dans le cadre des données censurées. Nous obtenons les lois asymptotiques, de l'estimateur et de la statistique de test proposés, lorsque le paramètre est un point frontière de son domaine.

Version française abrégée

Nous développons une procédure d'estimation semiparamétrique à deux étapes pour le paramètre d'association dans des modèles de copule dans le cadre des données censurées, suivant la procédure d'estimation proposée par [16]. Nous établissons les propriétés asymptotiques de l'estimateur du pseudo maximum de vraisemblance lorsque le paramètre appartient à la frontière de son domaine. Nous montrons que les lois limites classiques ne sont plus vérifiées. Nous proposons la statistique du rapport de vraisemblance généralisée pour tester l'indépendance. Nous obtenons la loi limite de la statistique proposée sous l'hypothèse nulle d'indépendance des marges. Ce travail constitue une extension des résultats de [3,2] au cas des données censurées.

1. Introduction and motivations

Many useful multivariate models for dependence between failure times T_1 and T_2 turn out to be generated by *parametric* families of copulas of the form $\{\mathbb{C}_\theta\colon\ \theta \in \boldsymbol{\Theta}\}$, typically indexed by a parameter $\theta \in \boldsymbol{\Theta} \subseteq \mathbb{R}$ (see, e.g., [10,11,7]). One advantage of copula models is that the margins are not specified and do not depend on the choice of the dependency structure, which allows to estimate the dependency and the margins separately. The reader may refer to the following books for excellent

E-mail addresses: salim.bouzebda@upmc.fr (S. Bouzebda), amor.keziou@upmc.fr (A. Keziou).

1631-073X/$ – see front matter © 2010 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.crma.2010.02.013

expositions of the basics of copula theory: [13] and [8]. In order to estimate the unknown *true* value of the parameter $\theta \in \boldsymbol{\Theta}$, which we denote, throughout the sequel, by $\theta_T \in \boldsymbol{\Theta}$, some semiparametric estimation procedures, based on the maximization, on the parameter space $\boldsymbol{\Theta}$, of properly chosen *pseudo-likelihood* criterion, have been proposed by [14] and studied by [6,16,18,17] among others. In each of these papers, some asymptotic normality properties are established for $\sqrt{n}(\hat{\theta}_n - \theta_T)$, where $\hat{\theta}_n$ denotes a properly chosen estimator of θ_T. This is achieved, provided that θ_T lies in the *interior*, denoted by $\mathring{\boldsymbol{\Theta}}$, of the parameter space $\boldsymbol{\Theta} \subseteq \mathbb{R}$. The case where $\theta_T \in \partial\boldsymbol{\Theta} := \overline{\boldsymbol{\Theta}} - \mathring{\boldsymbol{\Theta}}$ is a *boundary value* of $\boldsymbol{\Theta}$, has been studied in [2,3] in the case of complete data. However, the case of censored data has not been studied systematically until present when θ_T is a boundary value of the parameter space. Denote θ_0 the boundary value of the parameter and assume without loss of generality that the parameter space $\boldsymbol{\Theta}$ is of the form $\boldsymbol{\Theta} := [\theta_0, +\infty[$. Note that the case of the boundary value $\theta_T = \theta_0$ is very interesting since it corresponds to the hypothesis of independence of the margins for the majority of copulas models; see, e.g., [13] and [8]. Motivated by all this, we study the asymptotic properties of the maximum pseudo-likelihood estimate when $\theta_T = \theta_0$. We propose also a test of independence of margins based on the generalized pseudo-likelihood ratio statistic, and we give its limit law under the null hypothesis of independence. We show in particular that the limit laws of the estimate and the test statistic are not classical. The problems connected to this type of "non-regularity", for parametric models of densities with complete data, have been considered by several authors; see, e.g., [5,12,4,15,1].

The remainder of this Note is organized as follows. In the forthcoming section we present the estimation procedure, and we study the asymptotic properties of the estimate under the null hypothesis of independence. In Section 3, we give the limit law of the test statistic under independence. Section 4 reports some concluding remarks and possible developments. All proofs are postponed to Appendix A.

2. Estimation

Suppose that $\mathbb{C}_\theta$ is a distribution function with density c_θ on $(0,1)^2$ with respect to the Lebesgue measure for any $\theta \in \boldsymbol{\Theta}$. Let (T_1, T_2) denote the paired failure times, and (S_1, S_2) and (f_1, f_2) denote respectively the corresponding marginal survival functions and density functions. If (T_1, T_2) comes from $\mathbb{C}_{\theta_T}$ copula for some $\theta_T \in \boldsymbol{\Theta}$, then the joint survival function and density of (T_1, T_2) are given by

$$S(t_1, t_2) = \mathbb{C}_{\theta_T}\big(S_1(t_1), S_2(t_2)\big), \quad t_1, t_2 \geqslant 0, \qquad f(t_1, t_2) = c_{\theta_T}\big(S_1(t_1), S_2(t_2)\big) f_1(t_1) f_2(t_2), \quad t_1, t_2 \geqslant 0. \tag{1}$$

We recall the principle of the maximum pseudo-likelihood procedure studied by [16]. Let (C_1, C_2) denote paired censoring times. For $j = 1, \ldots, n$, $i = 1, 2$, assume that $(T_{1,j}, T_{2,j})$ and $(C_{1,j}, C_{2,j})$ are independent and random samples with continuous survival function S and G, respectively. For each j, we observe $X_{i,j} := T_{i,j} \wedge C_{i,j}$ and $\delta_{i,j} := \mathbb{1}_{\{X_{i,j} = T_{i,j}\}}$. We estimate S_1 and S_2 by the Kaplan–Meier estimators [9] denoted by $\hat{S}_{1,n}$ and $\hat{S}_{2,n}$. For $j = 1, \ldots, n$, write (u_j, v_j) for $(S_1(X_{1,j}), S_2(X_{2,j}))$. Then given (u_j, v_j), $j = 1, \ldots, n$, the likelihood of θ is

$$\prod_{j=1}^{n} L(\theta, u_j, v_j) = \prod_{j=1}^{n} c_\theta(u_j, v_j)^{\delta_{1,j}\delta_{2,j}} \frac{\partial \mathbb{C}_\theta(u_j, v_j)^{\delta_{1,j}(1-\delta_{2,j})}}{\partial u_j} \frac{\partial \mathbb{C}_\theta(u_j, v_j)^{\delta_{2,j}(1-\delta_{1,j})}}{\partial v_j} \mathbb{C}_\theta(u_j, v_j)^{(1-\delta_{1,j})(1-\delta_{2,j})}. \tag{2}$$

Let $\ell(\theta, u_j, v_j)$ denote the log of $L(\theta, u_j, v_j)$, and $U_\theta(\theta, u_j, v_j)$ the score function of θ, i.e., the derivative of log of (2) with respect to θ. The semiparametric maximum likelihood estimator $\hat{\theta}_n$ of θ_T is the solution of the estimating equation

$$U_\theta(\theta, \hat{S}_{1,n}, \hat{S}_{2,n}) := \sum_{j=1}^{n} \frac{\partial \ell(\theta, \hat{S}_{1,n}(X_{1,j}), \hat{S}_{2,n}(X_{2,j}))}{\partial \theta} = 0. \tag{3}$$

The following notations will be needed. Let

$$\begin{aligned}
&W_\theta(\theta, u, v) := \frac{\partial \ell(\theta, u, v)}{\partial \theta}, \qquad V_\theta(\theta, u, v) := \frac{\partial^2 \ell(\theta, u, v)}{\partial \theta^2},\\
&V_{\theta,1}(\theta, u, v) := \frac{\partial^2 \ell(\theta, u, v)}{\partial \theta \partial u}, \qquad V_{\theta,2}(\theta, u, v) := \frac{\partial^2 \ell(\theta, u, v)}{\partial \theta \partial v},\\
&t_{01} := \sup\big\{t\colon\ P(T_1 > t, C_1 > t) > 0\big\} \quad \text{and} \quad t_{02} := \sup\big\{t\colon\ P(T_2 > t, C_2 > t) > 0\big\}.
\end{aligned}$$

Letting

$$\begin{aligned}
\tau_1^2 &:= \mathbb{E}\big[-V_\theta\big(\theta_0, S_1(X_{1,1}), S_2(X_{1,2})\big)\big] = \int_A -V_\theta\big(\theta_T, S_1(t_1), S_2(t_2)\big)\,\mathrm{d}H_{\theta_T}(t_1, t_2, \delta_1, \delta_2),\\
\tau_2^2 &:= \mathbb{E}\big[\big\{I_1(X_{1,1}, \delta_{1,1}, \theta_T) + I_2(X_{1,2}, \delta_{1,2}, \theta_T)\big\}^2\big] = \int_A \big\{I_1(t_1, \delta_1, \theta_T) + I_2(t_2, \delta_2, \theta_T)\big\}^2\,\mathrm{d}H_{\theta_T}(t_1, t_2, \delta_1, \delta_2),
\end{aligned} \tag{4}$$

where H_{θ_T} is the joint distribution of $(X_{1,j},\delta_{1,j})$ and $(X_{2,j},\delta_{2,j})$, and $A:=[0,t_{01}]\times[0,t_{02}]$. For $j=1,\dots,n$, I_1 and I_2 are defined by

$$I_1(X_{1,j},\delta_{1,j},\theta_T):=\int_A V_{\theta,1}\big(\theta_T,S_1(t_1),S_2(t_2)\big)I_1^0(X_{1,j},\delta_{1,j})(t_1)\,\mathrm{d}H_{\theta_T}(t_1,t_2,\delta_1,\delta_2),$$

$$I_2(X_{2,j},\delta_{2,j},\theta_T):=\int_A V_{\theta,2}\big(\theta_T,S_1(t_1),S_2(t_2)\big)I_2^0(X_{2,j},\delta_{2,j})(t_2)\,\mathrm{d}H_{\theta_T}(t_1,t_2,\delta_1,\delta_2),$$

where

$$I_1^0(X_{1,j},\delta_{1,j})(t_1):=-S_1(t_1)\left[\int_0^{t_1}\frac{1}{P(T_1\geqslant u,C_1\geqslant u)}\,\mathrm{d}N_{1,j}(u)-\int_0^{t_1}\frac{\mathbb{1}_{\{X_{1,j}\geqslant u\}}}{P(T_1\geqslant u,C_1\geqslant u)}\,\mathrm{d}\Lambda_1(u)\right],$$

$$I_2^0(X_{2,j},\delta_{2,j})(t_2):=-S_2(t_2)\left[\int_0^{t_2}\frac{1}{P(T_2\geqslant u,C_2\geqslant u)}\,\mathrm{d}N_{2,j}(u)-\int_0^{t_2}\frac{\mathbb{1}_{\{X_{2,j}\geqslant u\}}}{P(T_2\geqslant u,C_2\geqslant u)}\,\mathrm{d}\Lambda_2(u)\right],$$

and

$$N_{i,j}(u):=\mathbb{1}_{\{X_{i,j}\leqslant u,\delta_{i,j}=1\}},\qquad \Lambda_i:=\log S_i,\quad i=1,2;\ j=1,\dots,n.$$

Under conditions (C.1)–(C.2) below, and when θ_T is an interior point of $\boldsymbol{\Theta}$, [16] show that, as $n\to\infty$,

$$\sqrt{n}(\hat\theta_n-\theta_T)\to N\big(0,\tau^2\big)\tag{5}$$

in distribution with variance $\tau^2:=(\tau_1^2+\tau_2^2)/\tau_1^4$. In the sequel, all derivatives of $\ell(\theta,\cdot,\cdot)$ are taken in the appropriate side. To describe the limiting behavior of $\hat\theta_n$, we will make use of the following conditions.

(C.1) Standard regularity conditions for the parametric maximum likelihood estimate;
(C.2) $W_\theta(\theta,S_1(t_1),S_2(t_2))$, $V_\theta(\theta,S_1(t_1),S_2(t_2))$, $V_{\theta,1}(\theta,S_1(t_1),S_2(t_2))$ and $V_{\theta,2}(\theta,S_1(t_1),S_2(t_2))$ are continuous and bounded for $(t_1,t_2)\in A$.

The asymptotic properties of $\hat\theta_n$ can then be summarized as follows.

Theorem 2.1. *Let the conditions* (C.1)–(C.2) *be fulfilled. Whenever $\theta_T=\theta_0$, is on the boundary of $\boldsymbol{\Theta}:=[\theta_0,\infty)$. Then, as $n\to\infty$, we have the convergence in distribution*

$$\sqrt{n}(\hat\theta_n-\theta_T)\xrightarrow{d}Z_+:=Z\mathbb{1}_{\{Z>0\}},\tag{6}$$

where $Z\overset{d}{:=}N(0,\sigma^2)$ denotes a centered normal random variable with variance $\sigma^2:=1/\tau_1^2$.

Remark 1. The asymptotic variance σ^2 in Theorem 2.1 may be consistently estimated by its empirical counterpart, as was done in [16, pp. 1389–1390]. Specifically, it may be obtained by replacing H_{θ_T} by its empirical distribution function H_n, and S_1, S_2 and θ_T by $\hat S_{1,n}$, $\hat S_{2,n}$ and $\hat\theta_n$.

Remark 2. When θ_T approaches the value corresponding to independence, i.e., $\theta_T\to\theta_0$, by integration by parts and by nothing that

$$\mathbb{E}\big(W_\theta(\theta_0,u,v)\partial\ell(\theta_0,u,v)/\partial u\big)=\mathbb{E}\big(W_\theta(\theta_0,u,v)\partial\ell(\theta_0,u,v)/\partial v\big)=0,$$

[16] showed that I_1 and I_2 converge to zero. By all this, at the independence τ_2^2 converges to zero, which implies that τ^2 tends to $1/\tau_1^2$, hence, $\hat\theta_n$ is asymptotically efficient estimate of θ_T when θ_T approaches θ_0.

3. Test of independence

In this section, we consider the independent test problem of margins in the previously considered parametric copula models. The null hypothesis to be tested is

$$\mathcal{H}_0\colon\ \mathbb{C}_{\theta_T}(u_1,u_2)=u_1u_2\quad\text{for all }u_1,u_2\in(0,1),$$

which is equivalent to $\mathcal{H}_0$: $\theta_T = \theta_0$ where θ_0 is the boundary value of the parameter space $\boldsymbol{\Theta}$. The alternative hypothesis, $\mathcal{H}_1$: $\theta_T \neq \theta_0$, is naturally composite. The corresponding generalized pseudo-likelihood ratio statistic is given by

$$\mathbf{T}_n := \mathbf{T}_n(\theta_0, \hat{\theta}_n) := 2\sum_{j=1}^{n} \ell\big(\hat{\theta}_n, \hat{S}_{1,n}(X_{1,j}), \hat{S}_{2,n}(X_{2,j})\big) - 2\sum_{j=1}^{n} \ell\big(\theta_0, \hat{S}_{1,n}(X_{1,j}), \hat{S}_{2,n}(X_{2,j})\big).$$

The following theorem gives the limiting law of the statistic $\mathbf{T}_n$ under $\mathcal{H}_0$.

Theorem 3.1. *Assume that the conditions of Theorem* 2.1 *hold. Then, under the null hypothesis $\mathcal{H}_0$, the statistic $\mathbf{T}_n$ converges in distribution, as $n \to \infty$, to the random variable $W^2 \mathbb{1}_{\{W>0\}}$, where $W \overset{d}{:=} N(0, 1)$ is a standard normal random variable.*

Remark 3. An application of Theorem 3.1, leads to reject the null hypothesis of independence $\mathcal{H}_0$: $\theta_T = \theta_0$, whenever the value of the statistic $\mathbf{T}_n$ exceeds $q_{1-\alpha}$, namely, the $(1-\alpha)$-quantile of the law of the random variable $W^2 \mathbb{1}_{\{W>0\}}$. The corresponding test is then, asymptotically of level α, when $n \to \infty$. The critical region is, accordingly, given by

$$CR := \{\mathbf{T}_n > q_{1-\alpha}\}.$$

4. Concluding remarks and possible developments

We have addressed the problem of testing the independence of margins in parametric copula models, with unknown and nonparametric margins, for censored data. For the majority of copula models, the value of the parameter corresponding to the null hypothesis of independence is a boundary value of the parameter space. We have derived the limit law of the semiparametric likelihood statistic under the null hypothesis; it is shown that the limit law of the generalized pseudo-likelihood ratio statistic is a mixture of chi-square law with one degree of freedom and Dirac measure at zero. A test of independence, based on this statistic, is then proposed. It would be interesting to study the asymptotic properties of the statistic under the alternative hypothesis and its optimality in some sense.

Acknowledgements

The authors are grateful to the referees for their useful suggestions and constructive criticisms on earlier drafts of this work.

Appendix A

Proof of Theorem 2.1. Using similar arguments as in [15], we can show that $\sqrt{n}(\hat{\theta}_n - \theta_T) = O_P(1)$ when θ_T is an interior or a boundary point of its domain $\boldsymbol{\Theta} := [\theta_0, \infty[$. At the independence, i.e., when $\theta_T = \theta_0$, by a Taylor expansion, we obtain for any θ satisfying $\theta - \theta_0 = O_P(1/\sqrt{n})$,

$$\frac{1}{n}\sum_{j=1}^{n} \ell\big(\theta, \hat{S}_{1,n}(X_{1,j}), \hat{S}_{2,n}(X_{2,j})\big) = \frac{1}{n}\sum_{j=1}^{n} \ell\big(\theta, S_1(X_{1,j}), S_2(X_{2,j})\big) + o_P(1/n), \tag{7}$$

and $\hat{\theta}_n - \bar{\theta}_n = o_P(1/\sqrt{n})$, where $\bar{\theta}_n$ is the parametric maximum likelihood, i.e.,

$$\bar{\theta}_n := \arg\max_{\theta \in \boldsymbol{\Theta}} \frac{1}{n}\sum_{j=1}^{n} \ell\big(\theta, S_1(X_{1,j}), S_2(X_{2,j})\big). \tag{8}$$

Furthermore, we can write by a Taylor expansion

$$\begin{aligned}&\frac{1}{n}\sum_{j=1}^{n} \ell\big(\theta, S_1(X_{1,j}), S_2(X_{2,j})\big) - \frac{1}{n}\sum_{j=1}^{n} \ell\big(\theta_0, S_1(X_{1,j}), S_2(X_{2,j})\big)\\ &\quad = -\tau_1^2\big(Z_n - (\theta - \theta_0)\big)^2 + n^{-2}\tau_1^{-2}U_\theta(\theta_0, S_1, S_2)^2 + O_P(1)|\theta - \theta_0|^3,\end{aligned}$$

where $Z_n := n^{-1}\tau_1^{-2}U_\theta(\theta_0, S_1, S_2)$. Since $\boldsymbol{\Theta} = [\theta_0, +\infty)$ is a convex set, making use of [15, Lemma 1], it holds that $\tilde{\theta}_n - \bar{\theta}_n = o_P(1/\sqrt{n})$, where

$$\tilde{\theta}_n := \arg\max_{\theta \in \boldsymbol{\Theta}} -\big(Z_n - (\theta - \theta_0)\big)^2 \tau_1^2. \tag{9}$$

Observe that the maximum of the quadratic positive function (9) is achieved at $\tilde{\theta}_n$ satisfying $\sqrt{n}(\tilde{\theta}_n - \theta_0) = \sqrt{n} Z_n \mathbb{1}_{\{\sqrt{n} Z_n > 0\}}$. Hence, the limit distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is the distribution of the random variable $Z \mathbb{1}_{\{Z>0\}}$ where $Z \overset{d}{:=} N(0, \sigma^2)$ denotes a centered normal random variable with variance $\sigma^2 := 1/\tau_1^2$. □

Proof of Theorem 3.1. As above, by a Taylor expansion, we can show that

$$\mathbf{T}_n(\theta_0, \hat{\theta}_n) = \sup_{\theta \in \boldsymbol{\Theta}} -\tau_1^2 \big(\sqrt{n} Z_n - \sqrt{n}(\theta - \theta_0)\big)^2 + o_P(1), \tag{10}$$

where $Z_n := n^{-1} \tau_1^{-2} U_\theta(\theta_0, S_1, S_2)$. Note that the supremum of the quadratic function in (10), on $\boldsymbol{\Theta} := [\theta_0, +\infty)$, is achieved at $\theta = \theta_0 + Z_n \mathbb{1}_{\{Z_n > 0\}}$. Hence, as n tens to infinity, the limit distribution of $\mathbf{T}_n(\theta_0, \hat{\theta}_n)$ is the distribution of $W^2 \mathbb{1}_{\{W>0\}}$, where W is a standard normal random variable. □

References

[1] D.W.K. Andrews, Estimation when a parameter is on a boundary, Econometrica 67 (6) (1999) 1341–1383.

[2] S. Bouzebda, A. Keziou, A test of independence in some copula models, Math. Methods Statist. 17 (2) (2008) 123–137.

[3] S. Bouzebda, A. Keziou, A new test procedure of independence in copula models via χ^2-divergence, Comm. Statist. Theory Methods 39 (1) (2010) 1–20.

[4] D. Chant, On asymptotic tests of composite hypotheses in nonstandard conditions, Biometrika 61 (1974) 291–298.

[5] H. Chernoff, On the distribution of the likelihood ratio, Ann. Math. Statist. 25 (1954) 573–578.

[6] C. Genest, K. Ghoudi, L.-P. Rivest, A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, Biometrika 82 (3) (1995) 543–552.

[7] H. Joe, Parametric families of multivariate distributions with given margins, J. Multivariate Anal. 46 (2) (1993) 262–282.

[8] H. Joe, Multivariate Models and Dependence Concepts, Monogr. Statist. Appl. Probab., vol. 73, Chapman & Hall, London, 1997.

[9] E.L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations, J. Amer. Statist. Assoc. 53 (1958) 457–481.

[10] G. Kimeldorf, A. Sampson, One-parameter families of bivariate distributions with fixed marginals, Comm. Statist. 4 (1975) 293–301.

[11] G. Kimeldorf, A. Sampson, Uniform representations of bivariate distributions, Comm. Statist. 4 (7) (1975) 617–627.

[12] P.A.P. Moran, Maximum-likelihood estimation in non-standard conditions, Proc. Cambridge Philos. Soc. 70 (1971) 441–450.

[13] R.B. Nelsen, An Introduction to Copulas, Lecture Notes in Statist., vol. 139, Springer-Verlag, New York, 1999.

[14] D. Oakes, Multivariate survival distributions, J. Nonparametr. Stat. 3 (3–4) (1994) 343–354.

[15] S.G. Self, K.-Y. Liang, Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, J. Amer. Statist. Assoc. 82 (398) (1987) 605–610.

[16] J.H. Shih, T.A. Louis, Inferences on the association parameter in copula models for bivariate survival data, Biometrics 51 (4) (1995) 1384–1399.

[17] H. Tsukahara, Semiparametric estimation in copula models, Canad. J. Statist. 33 (3) (2005) 357–375.

[18] W. Wang, A.A. Ding, On assessing the association for bivariate current status data, Biometrika 87 (4) (2000) 879–893.