Abstract

A test for the equality of marginal distributions of bi-dimensional distribution functions represented by a parametric copula and completely unknown marginals is proposed. To cite this article: V. Bagdonavičius et al., C. R. Acad. Sci. Paris, Ser. I 346 (2008). © 2008 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

Résumé


Version française abrégée

Soit (1) la fonction de survie $m$-dimensionelle, représentée en termes de la copule de survie $C$. On considère l’hypothèse (2) de l’égalité des fonctions de survie marginales contre l’alternative générale (3).

On suppose qu’on observe des vecteurs aléatoires indépendants bi-dimensionnels $(X_{1j}, X_{2j})$ de la fonction de survie (1), $j = 1, \ldots, n$.

Le test est basé sur les statistiques (4). On a démontré que sous les hypothèses $R, B, I$ et $H_0$ la loi limite de la statistique $\sqrt{n} (\hat{U}_1, \hat{U}_2)$ est normale et donc le test obtenu est un test asymptotique du type chi deux.

1. Introduction

Sklar [11] introduced a representation of $m$-dimensional distribution function as a composition of a distribution function concentrated in the unit cube $[0, 1]^m$ and marginal distribution functions (see Georges et al. [6]). The analo-
gous representation holds also for common survival function:
\[
\tilde{S}(x_1, \ldots, x_m) = C(S_1(x_1), \ldots, S_m(x_m); \alpha),
\]
where the function \(C\) is parametrized by the association parameter \(\alpha = (\alpha_1, \ldots, \alpha_r)\). The function \(C\) is called the survival copula (we call it simply a copula). The joint behavior of a random vector with continuous marginals \(F_i\) can be characterized uniquely by its associated copula.

Bagdonavičius et al. \cite{1} give a test when the copula is not specified. In some situations a specified copula with unknown marginals (Clayton copula, for example) is used as a good model (see Clayton and Cuzick \cite{4}). So we propose a test for the hypothesis
\[
H_0 : S_1 = S_2 = \cdots = S_m
\]
of the equality of the marginal distributions when the copula is specified and the marginals are completely unknown.

Denote by \(\lambda_i\) the hazard rate corresponding to \(S_i\) and let us consider the following wide class of alternatives: (cf. Bagdonavičius et al. \cite{1}): there exist \(\beta_i, \gamma_i, \sum_{i=2}^{m} (\beta_i^2 + \gamma_i^2) \neq 0\), such that
\[
H_0 : \lambda_i(t) = e^{\beta_i} \left\{ 1 + e^{\beta_i + \gamma_1} \Lambda(t) \right\}^{e^{-\gamma_1} - 1} \lambda(x) \quad (i = 2, \ldots, m),
\]
where \(\beta_2, \ldots, \beta_m, \gamma_2, \ldots, \gamma_m\) are unknown parameters, \(\lambda = \lambda_1\) is completely unknown, \(\Lambda = \Lambda_1\) is the cumulative hazard rate \(\Lambda(t) = \int_{0}^{t} \lambda(s) \, ds\).

The alternative contains a number of different possibilities: under it the hazard rates and the cumulative distribution functions may cross, the ratios of hazard rates may be constant, decrease or increase.

In this Note we consider the case of non-censored data. Unfortunately, there are essential complications to get asymptotic results in censored data case. There are implicit arguments that the methods will work only in the case of light censoring or under strong assumptions on copula parametric family.

Representation (1) can be used to estimate marginal distributions and principal dependence separately. Genest et al. \cite{5}, Shih and Louis \cite{10} give estimation procedures of the association parameter in the copula model with light censoring or under strong assumptions on copula parametric family.

2. The test

For simplicity consider the case \(m = 2\). Suppose that \(n\) independent two-dimensional vectors \((X_{1j}, X_{2j})\), each having joint survival function (1) are observed. Set
\[
p(y_1, y_2; \alpha) = \frac{\partial^2 C(y_1, y_2; \alpha)}{\partial y_1 \partial y_2}.
\]
Under the null hypothesis \(H_0\) and the alternative \(H_\theta\) the probability densities of \((X_{1j}, X_{2j})\) are
\[
\tilde{f}(x_1, x_2; 0, 0, \alpha) = p(S(x_1), S(x_2); \alpha) f(x_1) f(x_2)
\]
and
\[
\tilde{f}(x_1, x_2; \theta, \alpha) = p(S(x_1), \exp\{ -[1 - e^{\theta \lambda} \ln S(x_2)] e^{-\gamma} + 1\}; \alpha)
\times \exp\{ -[1 - e^{\theta \lambda} \ln S(x_2)] e^{-\gamma} + 1\} \left\{ 1 - e^{\theta \lambda} \ln S(x_2) \right\}^{e^{-\gamma} - 1} f(x_1) f(x_2),
\]
respectively; here \(S = e^{-\Lambda}, f = \lambda / S\). Set \((\theta, \alpha) = (\beta, \gamma, \alpha)\)
\[
h_1(v) = v \ln v, \quad q_1(v) = 1 + \ln v, \quad h_2(v) = v \left( 1 - \ln v \right) \ln(1 - \ln v) + \ln v, \quad q_2(v) = \ln v (1 - \ln(1 - \ln v)),
\]
\[
l(u, v, \alpha) = \ln p(u, v, \alpha), \quad g_i(u, v, \alpha) = l_i'(u, v, \alpha) h_i(v) + q_i(v), \quad i = 1, 2, \quad g_3(u, v, \alpha) = l'_{\alpha}(u, v, \alpha).
\]
For \(j = 1, 2, 3\) set
\[
U_j(S, \alpha) = \frac{1}{n} \sum_{i=1}^{n} g_j(S(X_{1i}), S(X_{2i}); \alpha).
\]
and for \( j = 1, 2 \) define
\[
\hat{U}_j = U_j(\hat{S}, \hat{\alpha}),
\]
where \( \hat{\alpha} \) verifies the equation \( U_3(\hat{S}, \alpha) = 0 \), \( \hat{S} = 1 - 2nF_{2n}/(2n + 1) \) with \( F_{2n} \) being the empirical distribution function from the unified sample. The random variable \( U_j(S, \alpha) \) can be written in the form
\[
U_j(S, \alpha) = n^{-1} \sum_{i=1}^{n} \int g_j(S(u_1), S(u_2), \alpha) \, dN_{1i}(u_1) \, dN_{2i}(u_2) = \int g_j(S(u_1), S(u_2), \alpha) \, F_n(du_1, du_2),
\]
where \( N_{1i}(x) = 1_{[X_{ij} \leq x]} \), \( F_n(u_1, u_2) = \frac{1}{n} \sum_{i=1}^{n} N_{1i}(u_1)N_{2i}(u_2) \) is the empirical distribution function. Under the standard regularity assumptions
\[
\begin{align*}
&\int_{0}^{\infty} \int_{0}^{\infty} \frac{\partial}{\partial \theta} \tilde{f}(x_1, x_2; \theta, \alpha) \, dx_1 \, dx_2 = \frac{\partial}{\partial \alpha} \int_{0}^{\infty} \int_{0}^{\infty} \tilde{f}(x_1, x_2; \theta, \alpha) \, dx_1 \, dx_2 = 0, \\
&\int_{0}^{1} \int_{0}^{1} \frac{\partial}{\partial \alpha} p(u_1, u_2; \alpha) \, du_1 \, du_2 = \frac{\partial}{\partial \alpha} \int_{0}^{1} \int_{0}^{1} p(u_1, u_2; \alpha) \, du_1 \, du_2 = 0,
\end{align*}
\]
with \( E|g_2(S(X_1), S(X_2); \alpha)| < \infty \), the random variables \( U_j \) have zero mean for \( j = 1, 2, 3 \) under \( H_0 \) and \( \alpha \) is true value of the parameter of association in copula. Analogously, for any function \( A : \mathbb{R} \to \mathbb{R} \) satisfying the conditions
\[
\int_{\mathbb{R}^2} \frac{\partial}{\partial \alpha} p(S(u_1), S(u_2), \alpha) A(v_1) \, dS(u_1) \, dS(v_2) = 0,
\]
such that \( E[A(X_{s1})g_2(F(X_{11}), F(X_{21}))] < \infty \) one can write that
\[
E(A(X_{s1})g_3(S(X_{11}), S(X_{12})) = E(A(X_{s1})I_{\alpha}(S(X_{11}), S(X_{21}))) = 0.
\]
The conditions (6) will be valid for any bounded continuous function \( A(\cdot) \) if
\[
\frac{1}{|\delta|} \int_{[0, 1]^2} \left| p(u_1, u_2, \alpha + \delta) - p(u_1, u_2, \alpha) - \delta p'_\alpha(u_1, u_2, \alpha) \right| \, du_1 \, du_2 \to 0,
\]
as \( \delta \to 0 \). In the same way the conditions
\[
\frac{1}{\|\delta\|^2} \int_{\mathbb{R}^2} \left| \tilde{f}(x_1, x_2; \theta + \delta, \alpha) - \tilde{f}(x_1, x_2; \theta, \alpha) - \delta \frac{\partial}{\partial \theta} \tilde{f}(x_1, x_2; \theta, \alpha) \right| \, dS(x_1) \, dS(x_2) \to 0
\]
as \( \delta \to 0 \) under \( \theta = 0 \) will imply that
\[
E(\mathbb{1}_{\{S(X_{11}) \geq u\}} g_2(S(X_{11}), S(X_{21}))) = 0 \quad \text{and} \quad E(\mathbb{1}_{\{S(X_{21}) \geq u\}} g_2(S(X_{11}), S(X_{21}))) = -h_1(u)
\]
for all \( u \in [0, 1] \). Essentially, for \( j = 1, 2 \),
\[
\sqrt{n} \hat{U}_j = \sqrt{n} U_j(\hat{S}, \hat{\alpha}) + (U_j)\alpha(S, \alpha) - \frac{1}{\sqrt{n}} \alpha + o_p(1)
\]
\[
= \frac{1}{\sqrt{n}} (U_j(\hat{S}, \hat{\alpha}) - (U_j)\alpha(S, \alpha) - ((U_3)\alpha^{-1})\alpha(S, \alpha) \cdot U_3(\hat{S}, \alpha) + o_p(1)
\]
\[
= \sqrt{n} \int g_j(S(x_1), \hat{S}(x_2); \alpha) \, + (U_j)\alpha(S, \alpha) \cdot I^{-1}(-\alpha) \cdot I'\alpha(\hat{S}(x_1), \hat{S}(x_2); \alpha) \, dF_n(x_1, x_2)
\]
\[
+ o_p(1),
\]
where \( I(\alpha) = -E((U_3)\alpha(S, \alpha) = E(U_3(S, \alpha))^2 \) is the information matrix in copula model with known marginals. In terms of the work by Berred and Malov [2] they are the type II multivariate rank statistics. Now we introduce the assumptions for asymptotic normality of the multivariate rank statistics:

**Assumption R.** Let the regularity assumption (5), (7) and (8) hold.
Assumption B. For \(i = 1, 2, 3\) and \(k = 1, 2,\)
\[
\left\| g_i(u_1, u_2; \theta) \right\| < M d^{a_1}(u_1) d^{a_2}(u_2), \quad \left\| \frac{\partial g_i(u_1, u_2; \theta)}{\partial u_k} \right\| < M_1 d^{a_1}(u_1) d^{a_2}(u_2)/d(u_k),
\]
(11)
where \(a_i = (-1/2 + \delta)/p_i\), with some constants \(M, \delta, p_i > 0, 1/p_1 + 1/p_2 = 1\), where \(d(u) = u(1 - u), u \in [0, 1]\).

Let the functions \(g_i(u_1, u_2; \alpha)\) be continuously differentiable in \(\alpha\) on \((0, 1)^2\), \(i = 1, 2, 3\), with
\[
g'_i(u_1, u_2; \theta) = \frac{\partial g_i(u_1, u_2; \alpha)}{\partial \alpha}.
\]
Recall also that \(g_3 \equiv l'_\alpha\). Let \(L'_0(\alpha) = \int_{(0,1)^2} l'_\alpha(u_1, u_2) C_0(du_1, du_2)\). If for \(i = 3\) the first of conditions in (11) holds in a neighborhood of \(\alpha\) then \(L'_0(\alpha)\) is also differentiable with respect to \(\alpha\) in this neighborhood and
\[
L''(\alpha) = \int_{(0,1)^2} g''_i(u_1, u_2; \alpha) C_0(du_1, du_2),
\]
where \(g''_{3\alpha} = \frac{\partial}{\partial \alpha} g_3\). It follows immediately from Lebesgue majoration theorem. The analogous property will be valid in the case of different marginal distributions too. Introduce the functions \(\omega_i(x_1, x_2; \delta), i = 1, 2, 3\), by
\[
\omega_i(x_1, x_2; \delta) = \sup_{\alpha: |\alpha - \alpha_0| < \delta} \left| g'_i(x_1, x_2; \alpha) - g'_i(x_1, x_2; \alpha_0) \right|
\]
for \(i = 1, 2\) and
\[
\omega_3(x_1, x_2; \delta) = \sup_{\alpha: |\alpha - \alpha_0| < \delta} \left| g''_3(x_1, x_2; \alpha) - g''_3(x_1, x_2; \alpha_0) \right| = \sup_{\alpha: |\alpha - \alpha_0| < \delta} \left| \frac{\partial^2 I(x_1, x_2; \alpha)}{\partial \alpha^2} - \frac{\partial^2 I(x_1, x_2; \alpha_0)}{\partial \alpha_0^2} \right|
\]
Assumption I. Let
\[
\int_{(0,1)^2} \frac{\partial^2}{\partial \alpha^2} p_\alpha(u_1, u_2) \, du_1 \, du_2 = 0,
\]
the information \(I(\alpha_0) > 0\), and
\[
g'_i(u_1, u_2; \alpha_0) \leq M d^{a_1}(u_1) d^{a_2}(u_2),
\]
where \(a_j = (-1 + \delta)/p_j\) for some positive constants \(\delta\) and \(p_j, j = 1, 2: p^{-1}_1 + p^{-1}_2 = 1\), \(i = 1, 2, 3\), and
\[
\omega_i(x_1, x_2; \delta) \leq M(\delta) d^{a_1}(x_1) d^{a_2}(x_2), \quad i = 1, 2, 3,
\]
for some constants \(M(\delta)\): \(\lim_{\delta \to 0} M(\delta) = 0\) and some \(b_k > -1/p_k, p^{-1}_1 + p^{-1}_2 = 1\) for \(k = 1, 2\).

The limit theorem for the statistic \((\hat{U}_1, \hat{U}_2)\) is following:

Theorem 2.1. Let the assumptions (R), (B) and (I) hold. Then under the hypothesis \(H_0\) the statistic \(\sqrt{n}(\hat{U}_1, \hat{U}_2)\) is asympototically normal, i.e.
\[
\sqrt{n}(\hat{U}_1, \hat{U}_2) \overset{d}{\to} N(0, \Sigma),
\]
where
\[
\Sigma = \text{Var}(G(S(X_{11}), S(X_{22}; \alpha) + (W_1(X_{11}, X_{21}, \alpha) + W_2(X_{11}, X_{21}, \alpha))/2),
\]
\[
W_k(X_{11}, X_{21}, \alpha) = \int_{(0,1)^2} \sum_{i=1}^{2} ((S(X_{i1}) \geq u_k) - (1 - u_k)) G_k(u_1, u_2; \alpha) C_0(du_1, du_2)
\]
with
\[ G(u_1, u_2; \alpha) = g(u_1, u_2; \alpha) + U'_\alpha(\alpha)g_3(u_1, u_2; \alpha)/I(\alpha), \quad G_k(u_1, u_2; \alpha) = \frac{\partial}{\partial u_k} G(u_1, u_2; \alpha), \]

\[ g(u_1, u_2; \alpha) = \left(g_1(u_1, u_2; \alpha), g_2(u_1, u_2; \alpha)\right)^T \quad \text{and} \quad U'(\alpha) = \left((U_1(S; \alpha))', (U_2(S; \alpha))'_\alpha\right)^T. \]

**Remark 1.** Under the conditions of the theorem by (7) and (9) the limit covariance matrix \( \Sigma \) has the following form \( \Sigma = G + C + W \), where

\[ G = \text{Var}(G(S(X_{11}), S(X_{21}); \alpha)), \quad W = \frac{1}{4} \text{Var}(W_1(X_{11}, X_{21}, \alpha) + W_2(X_{11}, X_{21}, \alpha)), \]

\[ C = -\frac{1}{2} \sum_{j=1}^2 \int_{(0,1)^2} h(u_j) S(X_{11}, \alpha)C_{\alpha}(du_1, du_2), \]

where \( h(u_j) = (h_1(u_j), h_2(u_j))^T \).

**Remark 2.** If the limit covariance matrix \( \Sigma \) is non-degenerated, and \( \hat{\Sigma} \) is a consistent estimator of the matrix \( \Sigma \), then under null hypothesis \( H_0 \) the limit distribution (as \( n \to \infty \)) of the test statistic

\[ Y^2 = (\sqrt{n} \tilde{U}_1, \sqrt{n} \tilde{U}_2) \hat{\Sigma}^{-1} (\sqrt{n} \tilde{U}_1, \sqrt{n} \tilde{U}_2)^T \]

is chi-square with two degrees of freedom, (see, for example, Greenwood and Nikulin [7]). The hypothesis \( H_0 \) is rejected with the approximate significance level \( \alpha, 0 < \alpha < 0.5 \), if \( Y^2 > \chi^2_{1-\alpha}(2) \).

The immediate variance estimator for \( \Sigma \) consists of two parts. To estimate the first term \( G \) of the covariance matrix \( \Sigma \) we use

\[ \hat{G} = \frac{1}{n} \sum_{i=1}^n \hat{G}(\hat{S}(X_{11}), \hat{S}(X_{21}); \hat{\alpha})^{\times 2} = \int_{\mathbb{R}^2} \hat{G}(\hat{S}(x_1), \hat{S}(x_2); \hat{\alpha})^{\times 2} dF_n(x_1, x_2), \]

where

\[ \hat{G}(u_1, u_2; \alpha) = g(u_1, u_2; \alpha) + U'_\alpha(\alpha)g_3(u_1, u_2; \alpha)/\hat{I}, \]

\[ U'_\alpha = \left((U_1)'_\alpha(\hat{\alpha}), (U_2)'_\alpha(\hat{\alpha})\right)^T \quad \text{and} \quad \hat{I} \text{ is either} \]

\[ \hat{I} = -L_{\alpha}(\hat{\alpha}) \quad \text{or} \quad \hat{I} = \tilde{I}_n = \int_{\mathbb{R}^m} (l'_{\alpha}(\hat{S}(x_1), \hat{S}(x_2)))^{\times 2} dF_n(x_1, x_2) = \frac{1}{n} \sum_{i=1}^n (l'_{\alpha}(\hat{S}(X_{11}), \hat{S}(X_{21})))^{\times 2}, \]

with \( A^{\times 2} = A A^T \) for a matrix \( A \). To estimate \( W = \text{Var}(W_1(X_1) + W_2(X_1)) \) we use the estimator \( \hat{W}_n = \hat{W}_n(\hat{S}; \hat{\alpha}) = \frac{1}{n} \sum_{i=1}^n \hat{W}_{n,i} \hat{W}_{n,i}^T \), where \( \hat{W}_{n,i} = \hat{W}_1(X_i) + \hat{W}_2(X_i) \) and

\[ \hat{W}_k(X_i) = \hat{W}_k(X_i; \hat{S}) = \int_{(0,1)^2} \sum_{j=1}^2 \left(1_{\hat{S}(x_{ij}) \geq u_k} - (1 - u_k)\right) G_k(u_1, u_2; \hat{\alpha}) dC_{\alpha}(u_1, u_2). \]

To estimate \( C \) we use the estimator

\[ \hat{C} = -\frac{1}{2} \sum_{j=1}^2 \int_{(0,1)^2} h(u_j) S(X_{11}, \alpha)C_{\alpha}(du_1, du_2). \]

Then the estimator of the covariance matrix \( \Sigma \) is \( \hat{\Sigma} = \hat{G} + \hat{W} + \hat{C} \). Set, for \( i = 1, 2, 3 \),

\[ \omega_i^2(x_1, x_2; \delta) = \sup_{\alpha:|\alpha - \alpha_0| < \delta} \left| g_i(x_1, x_2; \alpha) - g_i(x_1, x_2; \alpha_0) \right|. \]
Assumption V. Let

\[
\omega_i^\circ(x_1, x_2; \delta) \leq M(\delta) \prod_{k=1}^m b_1(x_1) b_2(x_2), \quad i = 1, 2, 3,
\]

for some constants \(M(\delta): \lim_{\delta \to 0} M(\delta) = 0\) and some \(b_k: b_k + a_k > -1/p_k, p_1^{-1} + p_2^{-1} = 1\) for \(k = 1, 2\), where \(a_k\) are the constants from (11). Suppose also that

\[
\omega_{\alpha_0}(\epsilon) = \int_{(0,1)^m} \sup_{|h| < \epsilon} \left( p_{\alpha_0 + h}(u)^{1/2} - p_{\alpha_0}(u)^{1/2} \right)^2 du \to 0 \quad \text{as} \quad \epsilon \to 0.
\]

Note that if to change the left-hand sides in (14) by \(g'_i(x_1, x_2; \alpha)\) and \(M(\delta)\) at the right-hand sides in (14) by some positive constant \(M\) we obtain more restrictive but easy verifiable assumptions.

**Theorem 2.2.** Under the conditions (R), (B), (I) and (V) the variance estimator \(\hat{\Sigma} = \hat{G} + \hat{W} + \hat{C}\) is consistent, i.e. \(\hat{\Sigma} \overset{P}{\longrightarrow} \Sigma\).

References