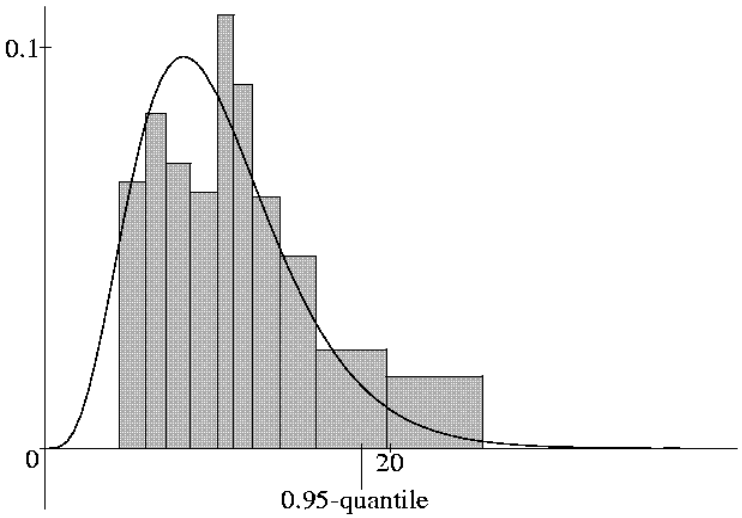
UNIVERSITÄT POTSDAM
Institut für Mathematik

A Note on:
Testing the Copula Based on Densities

Hannelore Liero



Mathematische Statistik und
Wahrscheinlichkeitstheorie

Universität Potsdam – Institut für Mathematik

Mathematische Statistik und Wahrscheinlichkeitstheorie

A Note on:
Testing the Copula Based on Densities

Hannelore Liero

Department of Mathematics, University of Potsdam, 14469 Potsdam, Germany

e-mail: liero@rz.uni-potsdam.de

Preprint 2006/02

November 2006

Impressum

Herausgeber: Mathematische Statistik und Wahrscheinlichkeitstheorie
am Institut für Mathematik

Adresse: Universität Potsdam
Am Neuen Palais 10
14469 Potsdam
Telefon:
Fax: +49-331-977 1500
E-mail: +49-331-977 1578
neisse@math.uni-potsdam.de

ISSN 1613-3307

A Note on Testing the Copula Based on Densities

Hannelore Liero
Institute of Mathematics, University of Potsdam
e-mail: liero@uni-potsdam.de

Abstract

We consider the problem of testing whether the density of a multivariate random variable can be expressed by a prespecified copula function and the marginal densities. The proposed test procedure is based on the asymptotic normality of the properly standardized integrated squared distance between a multivariate kernel density estimator and an estimator of its expectation under the hypothesis. The test of independence is a special case of this approach.

Keywords and phrases: Copula, multivariate kernel density estimator, asymptotic normality, independence

AMS subject classification: Primary 62 F 05

1 Introduction and Notation

Let X be a $\mathbb{R}^d$-valued random vector with distribution function H and density function h. The marginal distribution functions and the marginal densities are denoted by F_j and f_j, respectively. With $F(\underline{x}) := (F_1(x_1), \ldots, F_d(x_d))^T$, $\underline{x} = (x_1, \ldots, x_d)^T$ we can write

$$H(\underline{x}) \;=\; C(F(\underline{x})),$$

where C is the d-dimensional copula. Assume that C is differentiable with derivative

$$c(\underline{u}) \;=\; \frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \cdots \partial u_d} \qquad \underline{u} \in [0,1]^d.$$

Then the joint density h has the form

$$h(x_1, \ldots, x_d) \;=\; c(F_1(x_1), \ldots, F_d(x_d)) f_1(x_1) \cdots f_d(x_d).$$

At first we consider the simple hypothesis that the function c has a specific form, say c_0. That is we have to test

$$\mathcal{H}_0 : \quad c = c_0 \qquad \text{versus} \qquad \mathcal{H}_1 : \quad c \neq c_0. \tag{1}$$

Note that these are a nonparametric hypothesis and a nonparametric alternative, since the marginal distributions are not parameterized by a finite dimensional parameter. Furthermore, choosing $C_0(u_1, \ldots, u_d) \;=\; u_1 \cdots u_d$, and $c_0(\underline{u}) = 1$ we get the problem of testing independence, which is considered by several authors, for example Rosenblatt (1975) and Liero (2003).
The simple hypothesis can be extended to test whether the copula function belongs to a parametric class of copulas $\mathcal{C} \;=\; \{C(\cdot, \theta),\, \theta \in \Theta \subset \mathbb{R}^q\}$. For this problem Liebscher (2006) proposes goodness-of-fit tests based on estimates for the distribution function. In a further paper we will consider tests based on densities for this test problem.

Our test procedure is based on a weighted quadratic distance between a nonparametric kernel estimator for the density h and the smoothed hypothesis. Let $\underline{X}_1, \ldots, \underline{X}_n$ with $\underline{X}_i = (X_{1i}, \ldots, X_{di})^T$ be i.i.d. copies of $\underline{X}$. The kernel estimator for the density h is defined by

$$\hat{h}_n(\underline{x}) \;=\; \frac{1}{n b_n^d} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{1i}}{b_n}, \cdots, \frac{x_d - X_{di}}{b_n} \right) \;=:\; \frac{1}{n} \sum_{i=1}^{n} K_{b_n}(\underline{x} - \underline{X}_i),$$

where K is a kernel function mapping from $\mathbb{R}^d$ into $\mathbb{R}$, and b_n is a sequence of bandwidths tending to zero as n tends to infinity.
It is well-known that the kernel density estimator has a bias. To avoid this bias in the test procedure we will compare $\hat{h}_n$ not with the hypothetical density

$$c_0(F_1(x_1),\ldots,F_d(x_d))f_1(x_1)\cdots f_d(x_d) \;=\; c_0(F(\underline{x}))\prod_{j=1}^{d} f_j(x_j)$$

but with the estimated expectation of $\hat{h}_n$, where the expectation is taken under the null hypothesis. This expectation is given by

$$\mathsf{E}_0\hat{h}_n(\underline{x}) \;=\; \int K_{b_n}(\underline{x}-\underline{z})\,h(\underline{z})\,\mathrm{d}\underline{z} \;=\; \int K_{b_n}(\underline{x}-\underline{z})\,c_0(F(\underline{z}))\,f_1(z_1)\cdots f_d(z_d)\,\mathrm{d}\underline{z}.$$

To explain our estimation method assume for a moment that F is known. Then an unbiased estimator for the expectation $\mathsf{E}_0\hat{h}_n$ is

$$e_{n0}(\underline{x}) \;=\; \frac{(n-d)!}{n!}\sum_{\underline{i}} K_{b_n}(x_1-X_{1i_1},\ldots,x_d-X_{di_d})\,c_0(F(X_{1i_1},\ldots,X_{di_d}))$$

where the summation is taken over all vectors $\underline{i}=(i_1,\ldots,i_d)$ with $i_j\in\{1,\ldots,n\}$ and $i_j\neq i_{j'}$ for $j\neq j'$. Note that this estimator has the form of a U-statistic.
Of course, F is unknown - we replace it by its empirical version. Thus, finally, we estimate the hypothetical expectation by

$$\hat{e}_{n0}(\underline{x}) \;=\; \frac{(n-d)!}{n!}\sum_{\underline{i}} K_{b_n}(x_1-X_{1i_1},\ldots,x_d-X_{di_d})\,c_0(\hat{F}_{1n}(X_{1i_1}),\ldots,\hat{F}_{dn}(X_{di_d})),$$

where $\hat{F}_{jn}$ is the empirical marginal distribution of the j-th component.
As test statistic we propose the weighted integrated squared error

$$\hat{Q}_{n0} \;=\; \int \Big(\hat{h}_n(\underline{x}) \;-\; \hat{e}_{n0}(\underline{x})\Big)^2\, w(\underline{x})\,\mathrm{d}\underline{x},$$

where the weight function w is introduced to control the region of integration. It has to be chosen by the statistician. In the following section we will present a theorem stating the asymptotic normality of the standardized $\hat{Q}_{n0}$. Applying this theorem we get an asymptotic α-test by the rule: Reject $\mathcal{H}_0$, if

$$\hat{Q}_{n0} \;\geq\; \frac{z_\alpha\,\hat{\sigma}_{n0}}{nb_n^{d/2}} \;+\; \hat{\mu}_{n0}. \tag{2}$$

Here $\hat{\mu}_{n0}$ and $\hat{\sigma}^2_{n0}$ are suitable estimators for the standardizing terms in the limit theorem given below, and z_α is the $(1-\alpha)$-quantile of the standard normal distribution.

---

2 Limit theorem for the quadratic distance

Before we will state the limit theorem let us formulate the assumptions:

(A1) The density h of $\underline{X}$ has the form

$$h(x_1, \ldots, x_d) \;=\; c(F_1(x_1), \ldots, F_d(x_d)) f_1(x_1) \cdots f_d(x_d)$$

and is Lipschitz continuous in x.

(A2) The kernel K, $K : \mathbb{R}^d \to \mathbb{R}$, is a Lipschitz continuous density function with finite support.

(A3) The weight function w is nonnegative, piecewise continuous and bounded on $\mathbb{R}^d$

(A4) The bandwidth sequence satisfies:

$$b_n > 0, \qquad b_n \to 0 \qquad \text{and} \quad n b_n^d \to \infty.$$

Since we will present the limit statement not only under the null hypothesis set

$$Q_n \;=\; \int \Big(\hat{h}_n(\underline{x}) \;-\; e_n(\underline{x})\Big)^2 \, w(\underline{x}) \, \mathrm{d}\underline{x},$$

with

$$e_n(\underline{x}) \;=\; \frac{(n-d)!}{n!} \sum_{\underline{i}} K_{b_n}(x_1 - X_{1i_1}, \ldots, x_d - X_{di_d}) \, c(F(X_{1i_1}, \ldots, X_{di_d})),$$

and define

$$\mu_n \;=\; (n b_n^d)^{-1} \mu_{1n} \;-\; (n b_n)^{-1} \mu_{2n}$$

with

$$\mu_{1n} \;=\; \int \Omega_n(\underline{t}) \, w(\underline{t}) \, \mathrm{d}\underline{t} \qquad \Omega_n(\underline{t}) \;=\; \int K^2(\underline{x}) h(\underline{t} - \underline{x} b_n) \, \mathrm{d}\underline{x}$$

and

$$\mu_{2n} = \sum_{r=1}^{d} \int \Omega_{rn}(\underline{t})\, w(\underline{t})\, \mathrm{d}\underline{t}$$

where

$$\begin{aligned}
&\Omega_{rn}(\underline{t}) \\
&= \int K(\underline{x}) K(u_1, \ldots, u_{r-1}, x_r, u_{r+1}, \ldots, u_d) h(\underline{t} - \underline{x} b_n) \\
&\quad \times c(F(t_1 - u_1 b_n, \ldots, t_{r-1} - u_{r-1} b_n, t_r - x_r b_n, t_{r+1} - u_{r+1} b_n, \ldots, t_d - u_d b_n) \\
&\quad \times \prod_{\substack{j=1 \\ j \neq r}}^{d} f_j(t_j - u_j b_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_d\, \mathrm{d}u_1 \cdots \mathrm{d}u_{r-1}\, \mathrm{d}u_{r+1} \cdots \mathrm{d}u_d
\end{aligned}$$

and

$$\sigma^2 \;=\; 2 \int h^2(\underline{t})\, w^2(\underline{t})\, \mathrm{d}\underline{t} \int (\kappa^*(\underline{z}))^2\, \mathrm{d}\underline{z}, \qquad \kappa^*(\underline{z}) \;=\; \int K(\underline{u}) K(\underline{z} + \underline{u})\, \mathrm{d}\underline{u}.$$

Theorem 2.1 *Suppose that* (A1) - (A4) *are satisfied. Then*

$$\frac{n b_n^{d/2}}{\sigma} \left(Q_n \;-\; \mu_n \right) \xrightarrow{\mathcal{D}} \mathsf{N}(0, 1).$$

Since the empirical distribution functions $\hat{F}_{jn}$ are $\sqrt{n}$-consistent, the limit statement remains true if we replace e_n by

$$\hat{e}_n(\underline{x}) \;=\; \frac{(n-d)!}{n!} \sum_{\underline{i}} K_{b_n}(x_1 - X_{1i_1}, \ldots, x_d - X_{di_d})\, c(\hat{F}_{1n}(X_{1i_1}), \ldots, \hat{F}_{dn}(X_{di_d})).$$

For

$$\hat{Q}_n \;=\; \int \left(\hat{h}_n(\underline{t}) \;-\; \hat{e}_n(\underline{t}) \right)^2 w(\underline{t})\, \mathrm{d}\underline{t},$$

we obtain the following corollary:

Corollary 1 *Under the assumptions of Theorem* 2.1 *we have*

$$\frac{n b_n^{d/2}}{\sigma} \left(\hat{Q}_n \;-\; \mu_n \right) \xrightarrow{\mathcal{D}} \mathsf{N}(0, 1).$$

To apply this limit statement to the test problem we have to estimate the standardizing terms μ_n and σ^2. Note that

$$\Omega_n(\underline{t}) \;=\; \frac{1}{b_n^d}\int K^2\left(\frac{\underline{t}-\underline{x}}{b_n}\right)h(\underline{x})\,\mathrm{d}\underline{x} \;=\; \frac{1}{b_n^d}\mathsf{E}K^2\left(\frac{\underline{t}-\underline{X}_1}{b_n}\right),$$

thus an unbiased estimator for Ω_n is given by

$$\hat{\Omega}_n(\underline{t}) \;=\; \frac{1}{nb_n^d}\sum_{i=1}^{n} K^2\left(\frac{\underline{t}-\underline{X}_i}{b_n}\right).$$

For the term Ω_{rn} we obtain

$$\Omega_{rn}(\underline{t}) \;=\; \frac{1}{b_n^{2d-1}}\mathsf{E}K\left(\frac{\underline{t}-\underline{X}_r}{b_n}\right)K\left(\frac{t_1-X_{11}}{b_n},\cdots,\frac{t_d-X_{dd}}{b_n}\right)c(F(X_{11},\ldots,X_{dd}))$$

and an unbiased estimator is given by

$$\frac{(n-d)!}{n!b_n^{2d-1}}\sum_{\underline{i}} K\left(\frac{\underline{t}-\underline{X}_{i_r}}{b_n}\right)K\left(\frac{t_1-X_{1i_1}}{b_n},\cdots,\frac{t_d-X_{di_d}}{b_n}\right)c(F(X_{1i_1},\ldots,X_{di_d})),$$

so we estimate Ω_{rn} by

$$\begin{aligned}&\hat{\Omega}_{rn}(\underline{t})\\ &=\frac{(n-d)!}{n!b_n^{2d-1}}\sum_{\underline{i}} K\left(\frac{\underline{t}-\underline{X}_{i_r}}{b_n}\right)K\left(\frac{t_1-X_{1i_1}}{b_n},\cdots,\frac{t_d-X_{di_d}}{b_n}\right)c(\hat{F}_n(X_{1i_1},\ldots,X_{di_d})).\end{aligned}$$

So, finally we set

$$\hat{\mu}_n \;=\; (nb_n^d)^{-1}\int \hat{\Omega}_n(\underline{t})\,w(\underline{t})\,\mathrm{d}\underline{t} \;-\; (nb_n)^{-1}\sum_{r=1}^{d}\int \hat{\Omega}_{rn}(\underline{t})\,w(\underline{t})\,\mathrm{d}\underline{t}.$$

These estimators are $\sqrt{n}$-consistent. Since it is enough to estimate the variance consistently, we replace σ^2 by

$$\hat{\sigma}_n^2 \;=\; 2\int \hat{h}_n^2(\underline{t})\,w^2(\underline{t})\,\mathrm{d}\underline{t}\int(\kappa^*(\underline{z}))^2\,\mathrm{d}\underline{z}.$$

The statement that an asymptotic α-test is given by (2) is a consequence of the following corollary.

Corollary 2 *Under the assumptions of Theorem* 2.1 *we have*

$$\frac{nb_n^{d/2}}{\hat{\sigma}_n}\left(\hat{Q}_n \;-\; \hat{\mu}_n\right) \xrightarrow{\mathcal{D}} \mathsf{N}(0,1).$$

3 Testing independence of two random variables

Consider the case $d = 2$. The problem of testing independence of the components of the vector $\underline{X} = (X_1, X_2)$ is equivalent to the test problem (1) with $c_0(u_1, u_2) = 1$. Let us estimate the density h by a kernel density estimator with product kernel, i.e. we set

$$K(x_1, x_2) = K_1(x_1) \cdot K_2(x_2).$$

Applying the proposed procedure we obtain (with a slight modification) the following test: Reject the hypothesis of independence if

$$\hat{\mathrm{I}}_n \geq \frac{z_\alpha \, \hat{\sigma}_n^I}{n b_n} + \hat{\mu}_n^I \tag{3}$$

where

$$\hat{\mathrm{I}}_n = \int \left(\hat{h}_n(t_1, t_2) - \hat{f}_{1n}(t_1) \cdot \hat{f}_{2n}(t_2) \right)^2 w(t_1, t_2) \, \mathrm{d}t_1 \, \mathrm{d}t_2$$

and $\hat{f}_{1n}$ and $\hat{f}_{2n}$ are the kernel estimators of the marginal densities:

$$\hat{f}_{1n}(t_1) = \frac{1}{n b_n} \sum_{i=1}^{n} K_1 \left(\frac{t_1 - X_{1i}}{b_n} \right) \qquad \hat{f}_{2n}(t_2) = \frac{1}{n b_n} \sum_{i=1}^{n} K_2 \left(\frac{t_2 - X_{2i}}{b_n} \right).$$

The terms Ω_n^I and Ω_{rn}^I, $r = 1, 2$ are given by

$$\begin{aligned}
\Omega_n^I(t_1, t_2) &= \frac{1}{b_n^2} \int K_1^2 \left(\frac{t_1 - x_1}{b_n} \right) K_2^2 \left(\frac{t_2 - x_2}{b_n} \right) f_1(x_1) f_2(x_2) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \\
&= \frac{1}{b_n^2} \mathsf{E} K_1^2 \left(\frac{t_1 - X_{11}}{b_n} \right) K_2^2 \left(\frac{t_1 - X_{21}}{b_n} \right) \\
&= \frac{1}{b_n} \mathsf{E} K_1^2 \left(\frac{t_1 - X_{11}}{b_n} \right) \frac{1}{b_n} \mathsf{E} K_2^2 \left(\frac{t_1 - X_{21}}{b_n} \right),
\end{aligned}$$

$$\begin{aligned}
\Omega_{1n}^I(t_1, t_2) &= \frac{1}{b_n^2} \mathsf{E} K_1^2 \left(\frac{t_1 - X_{11}}{b_n} \right) K_2 \left(\frac{t_2 - X_{21}}{b_n} \right) \frac{1}{b_n} K_2 \left(\frac{t_2 - X_{22}}{b_n} \right) \\
&= \frac{1}{b_n} \mathsf{E} K_1^2 \left(\frac{t_1 - X_{11}}{b_n} \right) \left(\mathsf{E} \hat{f}_{2n}(t_2) \right)^2
\end{aligned}$$

and

$$\Omega^I_{2n}(t_1,t_2) = \frac{1}{b_n^2}\mathsf{E}K_1\left(\frac{t_1-X_{11}}{b_n}\right)K_2^2\left(\frac{t_2-X_{21}}{b_n}\right)\frac{1}{b_n}K_1\left(\frac{t_1-X_{12}}{b_n}\right)$$
$$= \frac{1}{b_n}\mathsf{E}K_2^2\left(\frac{t_2-X_{21}}{b_n}\right)\left(\mathsf{E}\hat{f}_{1n}(t_1)\right)^2$$

respectively. These terms can be estimated by

$$\hat{\Omega}^I_n(t_1,t_2) \;=\; \frac{1}{nb_n^2}\sum_{i=1}^n K_1^2\left(\frac{t_1-X_{1i}}{b_n}\right)K_2^2\left(\frac{t_2-X_{2i}}{b_n}\right).$$

and

$$\hat{\Omega}^I_{1n}(t_1,t_2) \;=\; \frac{1}{nb_n^2}\sum_{i=1}^n K_1^2\left(\frac{t_1-X_{1i}}{b_n}\right)K_2\left(\frac{t_2-X_{2i}}{b_n}\right)\hat{f}_{2n}(t_2)$$

$$\hat{\Omega}^I_{2n}(t_1,t_2) \;=\; \frac{1}{nb_n^2}\sum_{i=1}^n K_1\left(\frac{t_1-X_{1i}}{b_n}\right)K_2^2\left(\frac{t_2-X_{2i}}{b_n}\right)\hat{f}_{1n}(t_1)$$

leading finally to

$$\hat{\mu}^I_n = (nb_n^2)^{-1}\int \hat{\Omega}^I_n(t_1,t_2)\,w(t_1,t_2)\,\mathrm{d}t_1\,\mathrm{d}t_2$$
$$-\,(nb_n)^{-1}\int\left(\hat{\Omega}^I_{1n}(t_1,t_2)+\hat{\Omega}^I_{2n}(t_1,t_2)\right)\,w(t_1,t_2)\,\mathrm{d}t_1\,\mathrm{d}t_2.$$

The variance is estimated by

$$\hat{\sigma}^{I2}_n \;=\; 2\int \hat{f}^2_{1n}(t_1)\hat{f}^2_{2n}(t_2)\,w^2(t_1,t_2)\,\mathrm{d}t_1\,\mathrm{d}t_2\int(\kappa_1^*(z))^2\,\mathrm{d}z\int(\kappa_2^*(z))^2\,\mathrm{d}z.$$

Remark: A test for independence based on kernel densities was already considered by Rosenblatt (1975). But the approach given here differs from that proposed by Rosenblatt. He replaced the Ω_n's in the standardizing terms by their asymptotic expressions, then choosing the weight function $w(t_1,t_2) = (f_1(t_1)f_2(t_2))^{-1}$ the standardizing terms become independent of the unknown underlying marginal densities. The weight function in the quadratic distance I_n is then estimated by $(\hat{f}_{1n}\hat{f}_{2n})^{-1}$. Note that this approach requires stronger conditions on the smoothness of the densities and on the asymptotic behavior of the bandwidth sequence to ensure that the limit statement remains valid with these plug ins.

4 Proofs

The proof of Theorem 1 goes along the lines of the proof of the limit theorem for the integrated square error of multivariate nonparametric density estimators given by P. Hall (1984). The main difference is that we have here only the stochastic part of this deviation. For details of this approach see Liero (1999). Furthermore, in the model considered here we give a further term for the expectation of the integrated difference

$$\mathsf{E} \int \Big(\hat{h}_n(\underline{x}) \, - \, e_n(\underline{x}) \Big)^2 \, w(\underline{x}) \, \mathrm{d}\underline{x},$$

namely the term $(nb_n)^{-1}\mu_{2n}$.
The statements of the corollaries follows immediately, since the unknown terms are replaced by $\sqrt{n}$-consistent estimators. Since n^{-1} tends zero faster than the normalizing sequence in the limit statement $nb_n^{d/2}$ converges to infinity the estimation error vanishes.

References

1. Liero, H. Goodness of fit tests of L_2-type in: *Statistical Inference For Semiparametric Models and Applications,* Eds. M. Nikulin, N. Balakrishnan, N. Limnios and M. Mesbah, Birkhõuser, 2003

2. Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators , *Journal of Multivariate Analysis*, **14**, 1–16.

3. Liero, H. (1999). Global Measures of Deviation of Nonparametric Curve Estimators. *Habilitationsschrift, Mathematisch- Naturwissenschaftliche Fakultät der Universität Potsdam*

4. Liebscher, E. (2006) Goodness-of-fit tests in copula models, *Preprint University of Applied Science Merseburg, Germany.*