Memorandum 1866 (March 2008). ISSN 1874-4850. Available from: http://www.math.utwente.nl/publications
Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

Estimating dependence, model selection of copulas

Wilbert C.M. Kallenberg
Department of Applied Mathematics
Faculty of Electrical Engineering, Mathematics and Computer Science
University of Twente
P.O. Box 217, 7500 AE Enschede
The Netherlands

Abstract

Recently a new way of modeling dependence has been introduced considering a sequence of parametric copula models, covering more and more dependency aspects and approximating in this way the true copula density more and more. The method uses contamination families based on Legendre polynomials. It has been shown that in general after a few steps accurate approximations are obtained. In this paper selection of the adequate number of steps is treated, and estimation of the unknown parameters within the chosen contamination family is established. There should be a balance between the complexity of the model and the number of parameters to be estimated. High complexity gives a low model error, but a large stochastic or estimation error, while a very simple model gives a small stochastic error, but a large model error. Techniques from model selection are applied, thus letting the data tell us which aspects are important enough to capture into the model. Natural and simple estimators complete the procedure. Theoretical results show that the expected quadratic error is reduced by the selection rule to the same order of magnitude as in a classical parametric problem. The method is applied on a real data set, illustrating that the new method describes the data set very well: the error involved by the classical Gaussian copula is reduced with no fewer than 50%.

Keyword and phrases: copula, model selection, penalty function, Legendre polynomials, contamination family, nonlinear correlation.

2000 Mathematics Subject Classification: 62H12, 62H20, 62P05

1 Introduction

When modeling dependence with the multivariate normal distribution, all we need to know for the dependence structure are the linear correlations. Recently it is realized that considering only linear correlation is far too restricted and there is nowadays much attention for going beyond the linear dependence, in particular in finance and insurance (see e.g. Cherubini et al. (2004), Embrechts et al. (2002, 2003), McNeil et al. (2005)), but also in other areas like for instance hydrology (see e.g. Genest and Favre (2007)). For continuous multivariate distribution functions, the univariate margins and the multivariate dependence structure can be separated, using Sklar's (1959, 1996) theorem and the dependence structure can be represented by a so called copula. This copula is the multivariate distribution function of the random vector obtained by applying on each of the components its probability integral transformation, thus giving them uniform marginals. For a lot of results on copulas see also Joe (1997), Nelsen (1999), Cherubini et al. (2004), McNeil et al. (2005).

The great emphasis on copulas has been criticized by Mikosch (2006), leading to a passionate discussion, see e.g. Genest and Rémillard (2006). Some of the issues in that discussion concern the goodness-of-fit, the choice of the copulas and estimating aspects. In this paper this kind of issues is investigated following the new approach, started up in Kallenberg (2008).

A possible strategy to find a suitable copula is to consider a certain copula or family of copulas and to apply goodness-of-fit tests for testing the simple null hypothesis of a given copula, or the composite hypothesis of a parametric family of copulas, see e.g. Fermanian (2005), Panchenko (2005), Genest et al. (2008) and references therein. But in case of rejection, it is not clear what to do. In Biau and Wegkamp (2005) the problem of finding a particular copula, given a class of candidate copulas is attacked. They restrict attention to copulas with a bounded density. In their oracle inequality the upper bound consists of a model error term, expressing the distance between the true density and the parametric family of candidate copulas, and a second part giving the stochastic error or estimation error. Also in the approach presented here the total error is splitted up into the model error and the stochastic error, see (2.5) in Section 2.1 and (3.2) in Section 3.1. The modeling step has already been executed in Kallenberg (2008). The next steps, the required model selection and the estimation part are the topics of the present paper.

The modeling step, a new way of choosing a suitable copula to model dependence, can be described as follows. It consists of an intermediate approach between a parametric family (advantage: small estimation error, because only a few parameters have to be estimated; disadvantage: possibly a large model error due to a gap between the true distribution function and the chosen parametric family of copulas) and a nonparametric approach (advantage: no model error; disadvantage: large estimation error unless we have a huge number of observations). This is done by considering a sequence of parametric copula models, approximating the true copula more and more. Starting point is a given (family of) copula(s). It has been shown in Kallenberg (2008) that contamination families based on suitable Legendre polynomials give with a uniform start in general after a few steps accurate approximations, while applying a Gaussian start looks also very promising. A numerical study involving a large number of well known copulas illustrates clearly the results: errors of 10% when using the classical Gaussian copula are reduced in a few steps to only 1%. Even errors of more than 50% are reduced to only 4%, thus yielding for instance 0.030 as approximation for the true probability 0.029, see Table 11 in Kallenberg (2008).

Having established the nice accuracy of the approximations, we should select the dimension of the parametric contamination family and estimate the unknown parameters within the chosen contamination family. Considering higher and higher dimensions of the contamination families, we get in the limit the true density and in this way the method is "nonparametric". On the other hand, at every earlier step we have the advantage of a parametric family and hence we do not encounter a nonparametric estimation problem, but have to estimate (only) some parameters. There should be a balance between the complexity of the model and the number of parameters involved. To get such a balance we apply techniques from model selection, adapted to the problem at hand. In this way the data tell us which aspects are the most important ones to capture into our model. The model is kept as simple as possible, but if a more complicated model gives a much better fit, it is applied. The penalty in the selection step ensures that only a real improvement is awarded.

The unknown parameters within the chosen contamination family are estimated by moment estimators. They are linked up with the L^2-distance, which is also invoked for the modeling step. Therefore, these estimators are the natural (and simple!) estimators to apply. The whole procedure is easily implemented as is shown by a practical application.

Obviously, the marginal distributions should be estimated as well, but that is a very well known estimating problem with many solutions, depending on the assumptions made on the marginals (e.g. a parametric family). The comparison between the present approach based on

separating the marginal distributions and the copulas (including the influence of estimating the marginals on the modeling and estimation steps for the copulas) and on the other hand a fitting of the multivariate distribution as an entity to the data (see Mikosch (2006)) goes beyond the scope of this paper. To avoid too many technicalities we concentrate on bivariate distributions, but see Remark 2.7 in Kallenberg (2008) for an extension to the multivariate case.

The paper is organized as follows. In Section 2 the contamination families based on Legendre polynomials are introduced and the decomposition of the total error into the model error and the stochastic error is explained. Section 3 deals with the model selection problem. As we should avoid a large model error, the natural way to select the adequate model is to add new parameters as long as a substantial improvement of the model error is implied. A suitable penalty function, depending on the number of observations and the dimension of the model, does the job. Starting with a given copula or a given parametric family of copulas, with as typical examples the uniform density and the Gaussian copula densities, respectively, this leads to the estimated copula density in the appropriate contamination family, chosen by the data. Theoretical results are gathered in Section 4. It is shown that within a contamination family with a fixed, but unknown dimension, the selection rule selects the 'right' dimension with probability tending to 1 in a fast way. Moreover, also the right Fourier coefficients are chosen. The model error and stochastic error are investigated. It is shown that, although the approach has a nonparametric flavor, the selection rule reduces the expected quadratic error to the classical order $O(n^{-1})$ of a parametric problem. An application of the method to a real data set shows in detail the (easy) implementation of the method. Moreover, it illustrates the improvement due to taking nonlinear correlations into account. Although the linear correlation is prominently present in the data set, nevertheless an improvement of no fewer than 50% is obtained when going beyond the classical Gaussian copula.

2 Preliminaries

Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. random vectors with continuous distribution function $F_{X,Y}$. The marginal distribution functions of X and Y are denoted by F_X and F_Y, respectively. We consider the copulas (U_i, V_i) with

$$U_i = F_X(X_i), V_i = F_Y(Y_i), i = 1, ..., n.$$

In case of independence the simultaneous density of U_i and V_i equals 1 on the unit square, but due to dependence it may have another form. We assume that (U_i, V_i) has a density w.r.t. the Lebesgue measure on the unit square and we denote the true density of (U_i, V_i) by f and its distribution function by F. (Sometimes the distribution function of a copula is denoted by C and its density by c, but we prefer to use the notation F for the distribution function and f for its density.) Hence, we have the following relations $F_{X,Y}(x, y) = F(F_X(x), F_Y(y))$ and $F(u, v) = F_{X,Y}\left(F_X^{-1}(u), F_Y^{-1}(v)\right)$.

To model and estimate the true density f of (U_i, V_i), a sequence of parametric models is introduced, containing more and more dependency aspects. As starting point we take a given copula or a given parametric family of copulas, denoted as f_0. In particular we consider the independence start, given by the uniform density on the unit square, and the Gaussian copulas. The uniform density is denoted by $f_0^{un}(u, v)$ and the Gaussian copula densities by

$$f_0^G(u, v; \rho) = \frac{\varphi_2\left(\Phi^{-1}(u), \Phi^{-1}(v); \rho\right)}{\varphi\left(\Phi^{-1}(u)\right)\varphi\left(\Phi^{-1}(v)\right)}, \tag{2.1}$$

where Φ denotes the standard normal distribution function, φ its density and $\varphi_2(x, y; \rho)$ the density of the bivariate normal distribution with expectations 0, variances 1 and correlation coefficient ρ.

2.1 Contamination families

We start with copula density f_0. Let b_r be the r^{th} Legendre polynomial on $(0,1)$. The Legendre polynomials $b_0, ..., b_5$ are given by

$$\begin{aligned}
b_0(u) &= 1 \\
b_1(u) &= \sqrt{3}(2u-1) \\
b_2(u) &= \sqrt{5}(6u^2-6u+1) \\
b_3(u) &= \sqrt{7}(20u^3-30u^2+12u-1) \\
b_4(u) &= 3(70u^4-140u^3+90u^2-20u+1) \\
b_5(u) &= \sqrt{11}\left(252u^5-630u^4+560u^3-210u^2+30u-1\right).
\end{aligned}$$

We approximate $f - f_0$ by a linear combination of the functions $b_r(u)b_s(v)$, yielding the k-dimensional contamination family

$$f_k(u,v;\theta) = f_0(u,v) + \sum_{j=1}^{k} \theta_j b_{r_j}(u) b_{s_j}(v). \tag{2.2}$$

Denote the L_p-norm of f by $\|f\|_p$ and thus in particular the L_2-norm of f by

$$\|f\|_2 = \left\{ \int_0^1 \int_0^1 f(u,v)^2 du dv \right\}^{1/2}$$

and write the inner product of f and g as

$$< f, g >= \int_0^1 \int_0^1 f(u,v) g(u,v) du dv.$$

Let $f, f_0 \in L_2$, that is $\|f\|_2 < \infty, \|f_0\|_2 < \infty$. We then have (with equality in the L_2-sense)

$$f(u,v) - f_0(u,v) = \sum_{r,s} c_{rs} b_r(u) b_s(v)$$

with Fourier coefficients

$$\begin{aligned}
c_{rs} &=< f - f_0, b_r b_s >= \int\int \{f(u,v) - f_0(u,v)\} b_r(u) b_s(v) du dv \\
&= E_f b_r(U) b_s(V) - E_{f_0} b_r(U) b_s(V) = \rho\left(b_r(U), b_s(V); f\right) - \rho\left(b_r(U), b_s(V); f_0\right)
\end{aligned}$$

for copulas f and f_0, where $\rho\left(b_r(U), b_s(V); f\right)$ denotes the correlation coefficient of $b_r(U)$ and $b_s(V)$ under f. (Often we write simply E, var or P instead of E_f, var_f or P_f when the expectation, variance or probability under f is considered.) So, c_{rs} has a nice interpretation: it is just the change of the correlation coefficient of $b_r(U)$ and $b_s(V)$, when going from f_0 to f. For instance, c_{11} is just de change of the linear correlation coefficient of U and V, when going from f_0 to f.

Writing

$$f_k(u,v) = f_0(u,v) + \sum_{j=1}^{k} c_{r_j s_j} b_{r_j}(u) b_{s_j}(v)$$

the following proposition gives a "Pythagorean" result showing that f_k is the projection of f into the contamination family with "base" f_0: $\|f - f_k(\theta)\|_2^2$ is minimized by taking $\theta_j = c_{r_j s_j}$, because in that case the second term on the right-hand side of (2.3) vanishes.

Proposition 2.1 *For each $\theta \in \mathbb{R}^k$ we have*

$$\|f - f_k(\theta)\|_2^2 = \|f - f_k\|_2^2 + \|f_k - f_k(\theta)\|_2^2. \tag{2.3}$$

In particular, let $\widehat{c}_{rs}$ be an estimator of c_{rs} and let

$$\widehat{f}_k(u,v) = f_0(u,v) + \sum_{j=1}^{k} \widehat{c}_{r_j s_j} b_{r_j}(u) b_{s_j}(v), \tag{2.4}$$

then

$$\begin{aligned} \left\| f - \widehat{f}_k \right\|_2^2 &= \|f - f_k\|_2^2 + \left\| f_k - \widehat{f}_k \right\|_2^2 \\ &= \left(\sum_{r,s} c_{rs}^2 - \sum_{j=1}^{k} c_{r_j s_j}^2 \right) + \sum_{j=1}^{k} \left(c_{r_j s_j} - \widehat{c}_{r_j s_j} \right)^2. \end{aligned} \tag{2.5}$$

Proof. Because $f(u,v) - f_k(u,v) = \sum_{(r,s)\neq(r_j,s_j)} c_{rs} b_r(u) b_s(v)$ and $f_k(u,v) - f_k(u,v;\theta) = \sum_{j=1}^{k} \left(c_{r_j s_j} - \theta_j \right) b_{r_j}(u) b_{s_j}(v)$, orthonormality of the system $b_r(u)b_s(v)$ gives the result. ■

Equation (2.5) can be interpreted as

$$\text{Total Error} = \text{Model Error} + \text{Stochastic Error}.$$

2.2 Estimation

Obviously, c_{rs} depends on the unknown copula f and should be estimated. Writing

$$c_{rs} = \int\int b_r(u) b_s(v) dF(u,v) - E_{f_0} b_r(U) b_s(V)$$

the natural estimator of c_{rs} is obtained when replacing F by F_n, the empirical distribution function based on observations $(U_1, V_1), ..., (U_n, V_n)$; that is F_n gives probability mass n^{-1} to each of the points $(U_1, V_1), ..., (U_n, V_n)$. We then obtain the estimator

$$\widehat{c}_{rs} = \frac{1}{n} \sum_{i=1}^{n} b_r(U_i) b_s(V_i) - E_{f_0} b_r(U) b_s(V), \tag{2.6}$$

which also can be seen as the moment estimator of c_{rs}. So, indeed the moment estimators are the natural estimators in this context, linked up with the L_2-distance. When f_0 belongs to a parametric family, its parameter should be estimated as well. This aspect will be added in Section 3.3. Up to that point f_0 is assumed to be known.

3 Model selection, estimated density

Having established estimators of the parameters within the contamination family, the appropriate dimension of the contamination family should be chosen. Let m_n be a control sequence, giving the largest dimension for r, s under consideration with n observations. The estimated density then becomes

$$f_0(u,v) + \sum_{r=1}^{m_n} \sum_{s=1}^{m_n} \widehat{c}_{rs} b_r(u) b_s(v) \tag{3.1}$$

with $\widehat{c}_{rs}$ given by (2.6).

3.1 Model error, stochastic error

The model error and stochastic error when using this estimator is given in the following theorem. Before presenting and proving this theorem we give a lemma on the behavior of $\int b_r^4(u)du$.

Lemma 3.1 *For each $\varepsilon > 0$ we have*

$$\int |b_r(u)|^{4-\varepsilon}\, du = O(1) \text{ as } r \to \infty$$

and

$$\lim_{r\to\infty} \frac{\int b_r^4(u)du}{\log r} = \frac{6}{\pi^2}.$$

Proof. Essentially the result is given in Problem 91 on page 391 of Szegö (1959). A detailed analysis (not presented here) gives the limiting value $6/\pi^2$. ■

Theorem 3.2 *Let*

$$\begin{aligned} g(u,v) &= f(u,v) - f_0(u,v), \\ g_{m_n}(u,v) &= \sum_{r=1}^{m_n}\sum_{s=1}^{m_n} c_{rs} b_r(u) b_s(v) \\ \widehat{g}_{m_n}(u,v) &= \sum_{r=1}^{m_n}\sum_{s=1}^{m_n} \widehat{c}_{rs} b_r(u) b_s(v), \end{aligned}$$

then

$$\|g - \widehat{g}_{m_n}\|_2^2 = \|g - g_{m_n}\|_2^2 + \|g_{m_n} - \widehat{g}_{m_n}\|_2^2, \tag{3.2}$$

$$\begin{aligned} \|g - g_{m_n}\|_2^2 &= \sum_{r,s} c_{rs}^2 - \sum_{r=1}^{m_n}\sum_{s=1}^{m_n} c_{rs}^2, \\ \|g_{m_n} - \widehat{g}_{m_n}\|_2^2 &= \sum_{r=1}^{m_n}\sum_{s=1}^{m_n} (c_{rs} - \widehat{c}_{rs})^2 \end{aligned}$$

and

$$E\,\|g_{m_n} - \widehat{g}_{m_n}\|_2^2 = n^{-1}\sum_{r=1}^{m_n}\sum_{s=1}^{m_n} var(b_r(U)b_s(V)) = O\left(n^{-1}m_n^2 \log m_n\right) \text{ as } n \to \infty. \tag{3.3}$$

Moreover, if $f \epsilon L_{2+\varepsilon}$ for some $\varepsilon > 0$, we have

$$E\,\|g_{m_n} - \widehat{g}_{m_n}\|_2^2 = O\left(n^{-1}m_n^2\right) \text{ as } n \to \infty.$$

Proof. Because $g(u,v) - g_{m_n}(u,v) = \sum_{r>m_n \text{ and/or } s>m_n} c_{rs} b_r(u) b_s(v)$ and $g_{m_n}(u,v) - \widehat{g}_{m_n}(u,v) = \sum_{r=1}^{m_n}\sum_{s=1}^{m_n} (c_{rs} - \widehat{c}_{rs})\, b_r(u) b_s(v)$, orthonormality of the system $b_r(u)b_s(v)$ gives (3.2). Since $E\widehat{c}_{rs} = c_{rs}$, we have $E\,(c_{rs} - \widehat{c}_{rs})^2 = var(\widehat{c}_{rs}) = n^{-1} var(b_r(U)b_s(V))$. For all $f \in L_2$ and $1 \le r, s \le m_n$ application of the Cauchy Schwarz inequality gives

$$\begin{aligned} var(b_r(U)b_s(V)) &\le \int\int f(u,v) b_r^2(u) b_s^2(v) du dv \\ &\le \|f\|_2 \left\{ \int\int b_r^4(u) b_s^4(v) du dv \right\}^{1/2} \\ &\le \|f\|_2 \max_{1\le r\le m_n} \int b_r^4(u) du \end{aligned} \tag{3.4}$$

and (3.3) follows from Lemma 3.1. Let $f \epsilon L_{2+\varepsilon}$ for some $\varepsilon > 0$ and define $\eta = 2\varepsilon/(1+\varepsilon)$. By Hölder's inequality we get for all $1 \leq r, s \leq m_n$,

$$\begin{aligned} var(b_r(U)b_s(V)) &\leq \int\int f(u,v)b_r^2(u)b_s^2(v)dudv \\ &\leq \|f\|_{2+\varepsilon} \left\{ \int\int |b_r(u)|^{4-\eta} |b_s(v)|^{4-\eta} \, dudv \right\}^{(1+\varepsilon)/(2+\varepsilon)} \\ &\leq \|f\|_{2+\varepsilon} \left\{ \max_{1\leq r\leq m_n} \int |b_r(u)|^{4-\eta} \, du \right\}^{(2+2\varepsilon)/(2+\varepsilon)} \end{aligned} \tag{3.5}$$

and hence application of Lemma 3.1 gives $E\left\|g_{m_n} - \widehat{g}_{m_n}\right\|_2^2 = O\left(n^{-1}m_n^2\right)$, which completes the proof. ■

3.2 Selection rule

Taking all the coefficients $\widehat{c}_{rs}$ for $1 \leq r, s \leq m_n$ may be not a very good idea, because this introduces a large estimation error, in particular if a lot of the c_{rs} are small or even 0. Similarly as has been done in Kallenberg (2008) we consider only the largest Fourier coefficients and ignore the rest. Therefore, we replace the estimator from (3.1) by restricting to the k_n largest among $\widehat{c}_{rs}$ with $1 \leq r, s \leq m_n$, yielding

$$\widehat{f}\,(u,v) = f_0(u,v) + \sum_{j=1}^{k_n} \widehat{c}_{R_j S_j} b_{R_j}(u) b_{S_j}(v)$$

with

$$|\widehat{c}_{R_1 S_1}| \geq |\widehat{c}_{R_2 S_2}| \geq ... \geq \left|\widehat{c}_{R_{m_n^2} S_{m_n^2}}\right|.$$

(Note that due to continuity of the distributions we have strict inequalities here with probability 1 and hence with probability 1 the R_j, S_j are unique.) We have used here capitals R, S to stress that they are random variables, because they are not chosen in advance, but depending on the data. The obvious question then is: how large should we take k_n? If we take a larger k_n, the model error becomes smaller, but the stochastic error grows. The 'optimal' choice depends on f, but f is unknown. Therefore, a deterministic choice seems to be inadequate and hence we take a data driven selection of the dimension.

The idea behind choosing the appropriate dimension is that a higher dimension is only profitable if the approximation in the higher dimension gives a substantial improvement. The reason for it is that the higher dimension gives a more complicate description and moreover more parameters have to be estimated and hence a higher stochastic error occurs. Therefore, the idea is to base the selection rule on the improvement of the model error. The model error for $\widehat{f}_k$ from (2.4) is given by $\sum_{r,s} c_{rs}^2 - \sum_{j=1}^k c_{r_j s_j}^2$, cf. (2.5).

Hence, $\sum_{j=1}^k c_{r_j s_j}^2$ should grow sufficiently fast in order to take a higher dimension. For that purpose we introduce a penalty. The penalty is linear in k and decreasing in n. Obviously, we do not know the c_{rs} and therefore we replace them by $\widehat{c}_{rs}$. This leads to the following selection rule

$$S = \begin{cases} 0 & \text{if } \widehat{c}_{R_1 S_1}^2 < \Delta_n \\ \arg\max_{1\leq k\leq m_n^2} \left\{ \sum_{j=1}^k \widehat{c}_{R_j S_j}^2 - \Delta_n k \right\} & \text{otherwise} \end{cases}, \tag{3.6}$$

where Δ_n is decreasing and tends to 0 as $n \to \infty$. (Note that by continuity of the distributions the maximum is attained at a unique k with probability 1.) The selection rule can also be expressed as

$$S = \begin{cases} 0 & \text{if } \widehat{c}_{R_1 S_1}^2 < \Delta_n \\ \max\left\{ 1 \leq k \leq m_n^2 : \widehat{c}_{R_k S_k}^2 \geq \Delta_n \right\} & \text{otherwise} \end{cases}, \tag{3.7}$$

as is seen in the following lemma.

Lemma 3.3 *With probability* 1 *we have*

$$\arg \max_{1\le k\le m_n^2} \left\{ \sum_{j=1}^{k} \widehat{c}^2_{R_jS_j} - \Delta_n k \right\}$$
$$= \max \left\{ 1 \le k \le m_n^2 : \widehat{c}^2_{R_kS_k} \ge \Delta_n \right\}$$

if $\widehat{c}^2_{R_1S_1} \ge \Delta_n$.

Proof. Let

$$a_k = \sum_{j=1}^{k} \widehat{c}^2_{R_jS_j} - \Delta_n k.$$

With probability 1 the sequence $a_k - a_{k-1} = \widehat{c}^2_{R_kS_k} - \Delta_n$ is strictly decreasing and hence the sequence a_k is "strictly concave": it strictly grows as long as $\widehat{c}^2_{R_kS_k} - \Delta_n > 0$ and then it strictly goes down, implying that its maximum is attained at the (unique) largest k with $\widehat{c}^2_{R_kS_k} - \Delta_n \ge 0$, that is at $\max \left\{ 1 \le k \le m_n^2 : \widehat{c}^2_{R_kS_k} \ge \Delta_n \right\}$. ■

The expression in (3.7) has a nice interpretation: we proceed until the (estimated) Fourier coefficients are too small.

Classical penalties are for instance $n^{-1} \log n$ (in line with Schwarz's rule) or $2n^{-1}$ (in line with Akaike's criterion). However, we should realize that we are not going through the dimensions in an in advanced prescribed way, but we take so to say the "best" route by taking the largest (estimated) Fourier coefficients first. Obviously, we therefore should take a larger penalty. We can see this also from the formulation given by (3.7). Suppose that $c_{rs} = 0$ for all $r \ge K$ and/or $s \ge K$. Then we should have $S \le K^2$ with high probability. But $\widehat{c}^2_{R_{K^2+1}S_{K^2+1}}$ still may be relatively large, because it is the largest among $m_n^2 - K^2$ pairs $\widehat{c}^2_{rs}$. Therefore, it may be better to take a larger penalty, taking into account the variance of $\widehat{c}^2_{rs}$. In view of (3.4), noting that $\max_{1\le r\le m_n} \int b_r^4(u)du$ behaves like $6\pi^{-2} \log m_n$, one may think on penalties like

$$\Delta_n = n^{-1} (\log n) (\log m_n). \tag{3.8}$$

3.3 Estimated density

The estimated density now becomes

$$\widehat{f}(u,v) = f_0(u,v) + \sum_{j=1}^{S} \widehat{c}_{R_jS_j} b_{R_j}(u) b_{S_j}(v).$$

When f_0 belongs to a parametric family, $f_0(u,v;\tau)$, say, then the parameter τ should also be estimated. Note that $\widehat{c}_{rs}$ depends in that case on τ as well, see (2.6), and hence we end up with

$$\widehat{f}(u,v) = f_0(u,v;\widehat{\tau}) + \sum_{j=1}^{S} \widehat{c}_{R_jS_j}(\widehat{\tau}) b_{R_j}(u) b_{S_j}(v), \tag{3.9}$$

where

$$\widehat{c}_{rs}(\widehat{\tau}) = \frac{1}{n} \sum_{i=1}^{n} b_r(U_i) b_s(V_i) - \iint b_r(u) b_s(v) f_0(u,v;\widehat{\tau}) du dv.$$

When τ is given, we often use the shorter notation $\widehat{c}_{rs}$ instead of $\widehat{c}_{rs}(\tau)$, as we have done before, but this shorter notation $\widehat{c}_{rs}$ is never used when τ is estimated as well.

Remark 3.1 The growth of $\sum_{j=1}^{k} c_{r_j s_j}^2$ can also be measured relative to $\sum_{r,s} c_{rs}^2$. That is considering $\sum_{j=1}^{k} c_{r_j s_j}^2 - \Delta_n k \sum_{r,s} c_{rs}^2$ in the selection rule. On the one hand, as $\sum_{r,s} c_{rs}^2 < \infty$, the order of the penalty does not change in this variant. On the other hand, we should estimate in that case $\sum_{r,s} c_{rs}^2$, for instance by $\sum_{r=1}^{m_n} \sum_{s=1}^{m_n} \widehat{c}_{rs}^2$. However, it should be realized that $E\left(\widehat{c}_{rs}^2\right)$ does not tend to 0 if $c_{rs} \to 0$. For instance, take $f = f_0^{un}$, then $c_{11} = 0$, but $E\left(\widehat{c}_{11}^2\right) = n^{-1} \int\int \{b_1(u) b_1(v)\}^2 \, du dv = n^{-1} > 0$. Therefore, $\sum_{r=1}^{m_n} \sum_{s=1}^{m_n} \widehat{c}_{rs}^2$ may give too much noise and hence we prefer the more simple rule, given in (3.6).

Remark 3.2 In similar situations in testing theory selection rules like Schwarz's rule are applied. Starting with Ledwina (1994) a lot of papers on data-driven tests using (modifications of) Schwarz's selection rule have been appeared. Also the problem of testing independence is covered, see e.g. Kallenberg and Ledwina (1999). Occasionally, for the goodness-of-fit problem a density estimate fitted to the data in the dimension given by Schwarz's selection rule is applied, see e.g. Kallenberg and Ledwina (1997), Figure 5. However, as argued in Section 4.2 in Kallenberg (2008) the testing problem really differs from the estimation problem. The focus in testing is more on what is happening in the neighborhood of the null hypothesis, while in estimation all types of dependence should be dealt with. As a consequence the selection rule proposed here is different from the one in testing theory. In the testing problem the loglikelihoodratio or the score function comes in. The score statistic in the contamination model $f_k(u, v; \theta) = f_0(u, v) + \sum_{j=1}^{k} \theta_j b_{r_j}(u) b_{s_j}(v)$ becomes $n Z^\top I^{-1} Z$ with $Z = (Z_1, ..., Z_k)$,

$$\begin{aligned} Z_j &= n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \log f_k(U_i, V_i; 0) \\ &= n^{-1} \sum_{i=1}^{n} \frac{b_{r_j}(U_i) b_{s_j}(V_i)}{f_0(U_i, V_i)}, \end{aligned}$$

$^\top$ denoting the transpose and I the information matrix at $\theta = 0$. When f_0 is the uniform distribution, we get $Z_j = \widehat{c}_{r_j s_j}$ and I then equals the identity matrix, yielding $Z^\top I^{-1} Z = \sum_{j=1}^{k} \widehat{c}_{r_j s_j}^2$. However, for other densities f_0 we do not get this form. While in testing theory the natural approach is to consider the ratio f/f_0 (and hence a multiplicative structure), in estimation theory it seems more natural in connection with the L_2-norm to consider the additive contamination family as introduced in (2.2).

Remark 3.3 The problem of choosing the appropriate dimension is certainly not replaced by the problem of choosing m_n. In most practical situations there is not that much change for larger m_n. This can be seen from the extensive numerical results in Kallenberg (2008).

4 Theoretical results

In this section we investigate the asymptotic behavior of the new estimator $\widehat{f}$ presented in (3.9). Estimation of τ will be done by standard estimators for the parametric family, while no estimation at all is needed when f_0 is a given density. Therefore, we concentrate on the second part of (3.9) with $\widehat{c}_{R_j S_j}(\widehat{\tau})$ replaced by $\widehat{c}_{R_j S_j}(\tau)$, which often for simplicity is written as $\widehat{c}_{R_j S_j}$.

It has been seen in Kallenberg (2008) that already after a few steps the contamination model gives in general accurate approximations. However, the number of steps is not known beforehand and moreover, among the, for instance 5, largest Fourier coefficients we may find c_{63} or c_{77} and not necessarily only c_{rs} with the smallest r, s. Therefore, we consider the following contamination family as our true density

$$f(u, v) = f_0(u, v; \tau) + \sum_{r=1}^{K} \sum_{s=1}^{K} c_{rs} b_r(u) b_s(v) \tag{4.1}$$

with K fixed, but unknown. In other words, $c_{rs} = 0$ for $r \geq K+1$ and/or $s \geq K+1$. (Note that also for $r, s \leq K$ some of the c_{rs} can be 0.) This reflects that very small Fourier coefficients are unimportant. In practice, they can be considered as 0. For theoretical convenience it is more appropriate to really make them equal to 0. However, from which point on is not known beforehand.

The ordered $|c_{rs}| = \left|\int\int b_r(u)b_s(v)f(u,v)dudv - E_{f_0}b_r(U)b_s(V)\right|, 1 \leq r, s \leq m_n$ are denoted as

$$\begin{aligned} &\left|c_{r_1^* s_1^*}\right| \geq \left|c_{r_2^* s_2^*}\right| \geq ... \geq \left|c_{r_L^* s_L^*}\right| > 0 \\ &= \left|c_{r_{L+1}^* s_{L+1}^*}\right| = ... = \left|c_{r_{K^2}^* s_{K^2}^*}\right| = \left|c_{r_{K^2+1}^* s_{K^2+1}^*}\right| = ... \left|c_{r_{m_n^2}^* s_{m_n^2}^*}\right| \end{aligned} \tag{4.2}$$

with $L = 0$ if $c_{rs} = 0$ for all r, s (and thus $f = f_0$). So, $c_{r_L^* s_L^*}$ has to be interpreted as the last important Fourier coefficient.

4.1 Selecting the right dimension

We start with showing that S selects the 'right' dimension L with probability tending to 1, often at a fast rate.

Theorem 4.1 *Let f be given by (4.1) and L by (4.2). Let*

$$v_n = \max_{1 \leq r,s \leq m_n} var(b_r(U)b_s(V)), w_n = \max_{1 \leq r \leq m_n} \int b_r^4(u)du. \tag{4.3}$$

Suppose that

$$\lim_{n\to\infty} \Delta_n = 0, \lim_{n\to\infty} m_n = \infty, \lim_{n\to\infty} \frac{m_n\sqrt{\Delta_n}}{v_n} = 0, \lim_{n\to\infty} \frac{n\Delta_n}{\log m_n} = \infty. \tag{4.4}$$

Then, for each $\varepsilon > 0$,

$$P(S = L) \geq 1 - 2m_n^2 \exp\left(-\frac{n\Delta_n}{(2+\varepsilon)\, v_n}\right) \geq 1 - 2m_n^2 \exp\left(-\frac{n\Delta_n}{(2+\varepsilon)\, \|f\|_2\, w_n}\right) \tag{4.5}$$

for all $n \geq n(\varepsilon)$ and hence, assuming that

$$\lim_{n\to\infty} \Delta_n = 0, \lim_{n\to\infty} m_n = \infty, \lim_{n\to\infty} \frac{m_n\sqrt{\Delta_n}}{v_n} = 0, \lim_{n\to\infty} \frac{n\Delta_n}{(\log m_n)^2} = \infty, \tag{4.6}$$

we get

$$\lim_{n\to\infty} P(S = L) = 1.$$

If we replace

$$\lim_{n\to\infty} \frac{n\Delta_n}{(\log m_n)^2} = \infty$$

in (4.6) by

$$n\Delta_n \geq \log n(\log m_n)^2,$$

then for every $c > 0$,

$$P(S \neq L) = O(n^{-c}) \text{ as } n \to \infty. \tag{4.7}$$

If $f_0 \in L_{2+\delta}$ for some $\delta > 0$, then conditions (4.4) are sufficient for getting $\lim_{n\to\infty} P(S = L) = 1$ and replacing

$$\lim_{n\to\infty} \frac{n\Delta_n}{\log m_n} = \infty$$

in (4.4) by

$$n\Delta_n \geq \log n(\log m_n),$$

yields for every $c > 0$,

$$P(S \neq L) = O(n^{-c}) \text{ as } n \to \infty.$$

Before proving Theorem 4.1 we present some lemmas.

Lemma 4.2 *Let f be given by (4.1) and L by (4.2). Suppose that $\lim_{n\to\infty}\Delta_n = 0$. Then, for $L \geq 1$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log P\left(\widehat{c}^2_{R_L S_L} < \Delta_n\right) < 0.$$

Proof. Because

$$P\left(\widehat{c}^2_{R_L S_L} < \Delta_n\right) \leq P\left(\cup_{j=1}^{L}\left\{\widehat{c}^2_{r_j^* s_j^*} < \Delta_n\right\}\right) \leq \sum_{j=1}^{L} P\left(\widehat{c}^2_{r_j^* s_j^*} < \Delta_n\right)$$

and L is fixed, it suffices to prove that $P\left(\widehat{c}^2_{r_j^* s_j^*} < \Delta_n\right)$ is exponentially small for each $1 \leq j \leq L$. We have

$$\begin{aligned} P\left(\widehat{c}^2_{r_j^* s_j^*} < \Delta_n\right) &= P\left(\left|\widehat{c}_{r_j^* s_j^*}\right| < \sqrt{\Delta_n}\right) \\ &\leq P\left(\left|\widehat{c}_{r_j^* s_j^*} - E\widehat{c}_{r_j^* s_j^*}\right| > \left|c_{r_j^* s_j^*}\right| - \sqrt{\Delta_n}\right). \end{aligned}$$

Since $\lim_{n\to\infty}\left(\left|c_{r_j^* s_j^*}\right| - \sqrt{\Delta_n}\right) = \left|c_{r_j^* s_j^*}\right| > 0$, application of Chernoff's theorem shows that indeed $P\left(\widehat{c}^2_{r_j^* s_j^*} < \Delta_n\right)$ is exponentially small, thus completing the proof. ■

Lemma 4.3 *Let f be given by (4.1) and L by (4.2). Then*

$$P\left(\widehat{c}^2_{R_{L+1} S_{L+1}} \geq \Delta_n\right) \leq 2m_n^2 \exp\left\{-\frac{n\Delta_n}{2v_n + \frac{2}{3}(2m_n+1)\sqrt{\Delta_n}}\right\},$$

where v_n is given by (4.3).

Proof. We have

$$P\left(\widehat{c}^2_{R_{L+1} S_{L+1}} \geq \Delta_n\right) \leq P\left(\cup_{j=L+1}^{m_n^2}\left\{\widehat{c}^2_{r_j^* s_j^*} \geq \Delta_n\right\}\right) \leq \sum_{j=L+1}^{m_n^2} P\left(\widehat{c}^2_{r_j^* s_j^*} \geq \Delta_n\right).$$

Noting that, cf. Sansone (1959) p. 181,

$$\max_{0\leq u\leq 1} |b_r(u)| = \sqrt{2r+1}, \tag{4.8}$$

and that $E\widehat{c}_{r_j^* s_j^*} = c_{r_j^* s_j^*} = 0$ for $L+1 \leq j \leq m_n^2$, Bernstein's inequality (see e.g. Serfling (1980), p. 95) gives

$$\begin{aligned} &P\left(\left|\widehat{c}_{r_j^* s_j^*} - E\widehat{c}_{r_j^* s_j^*}\right| \geq \sqrt{\Delta_n}\right) \qquad (4.9) \\ &\leq 2\exp\left\{-\frac{n\Delta_n}{2var\left(b_{r_j^*}(U) b_{s_j^*}(V)\right) + \frac{2}{3}\sqrt{\left(2r_j^*+1\right)\left(2s_j^*+1\right)\Delta_n}}\right\} \\ &\leq 2\exp\left\{-\frac{n\Delta_n}{2v_n + \frac{2}{3}(2m_n+1)\sqrt{\Delta_n}}\right\}. \end{aligned}$$

Hence,

$$P\left(\widehat{c}^2_{R_{L+1}S_{L+1}} \geq \Delta_n\right) \leq \sum_{j=L+1}^{m_n^2} 2\exp\left\{-\frac{n\Delta_n}{2v_n + \frac{2}{3}(2m_n+1)\sqrt{\Delta_n}}\right\}$$
$$= 2\left(m_n^2 - L\right)\exp\left\{-\frac{n\Delta_n}{2v_n + \frac{2}{3}(2m_n+1)\sqrt{\Delta_n}}\right\}$$
$$\leq 2m_n^2\exp\left\{-\frac{n\Delta_n}{2v_n + \frac{2}{3}(2m_n+1)\sqrt{\Delta_n}}\right\},$$

as was to be proved. ■

Proof of Theorem 4.1. By definition (3.7) of S it follows that

$$P(S=L) = P\left(\widehat{c}^2_{R_LS_L} \geq \Delta_n, \widehat{c}^2_{R_{L+1}S_{L+1}} < \Delta_n\right)$$
$$= 1 - P\left(\widehat{c}^2_{R_LS_L} < \Delta_n \cup \widehat{c}^2_{R_{L+1}S_{L+1}} \geq \Delta_n\right)$$
$$= 1 - P\left(\widehat{c}^2_{R_LS_L} < \Delta_n\right) - P\left(\widehat{c}^2_{R_{L+1}S_{L+1}} \geq \Delta_n\right)$$

for $L \geq 1$ and $P(S=0) = P\left(\widehat{c}^2_{R_1S_1} < \Delta_n\right)$. Application of Lemma 4.2 gives $\limsup_{n\to\infty} \frac{1}{n}\log P\left(\widehat{c}^2_{R_LS_L} < \Delta_n\right) < 0$ for $L \geq 1$ and therefore

$$P\left(\widehat{c}^2_{R_LS_L} < \Delta_n\right) \leq \exp(-nc_1)$$

for some constant $c_1 > 0$ and all $n \geq n_1$. By Lemma 4.3 we get

$$P\left(\widehat{c}^2_{R_{L+1}S_{L+1}} \geq \Delta_n\right) \leq 2m_n^2\exp\left\{-\frac{n\Delta_n}{2v_n + \frac{2}{3}(2m_n+1)\sqrt{\Delta_n}}\right\}.$$

Let $\varepsilon > 0$. To get the first inequality in (4.5) it remains to be shown that for $n \geq n(\varepsilon)$,

$$\exp(-nc_1) + 2m_n^2\exp\left\{-\frac{n\Delta_n}{2v_n + \frac{2}{3}(2m_n+1)\sqrt{\Delta_n}}\right\} \tag{4.10}$$
$$\leq 2m_n^2\exp\left(-\frac{n\Delta_n}{(2+\varepsilon)v_n}\right),$$

or, equivalently,

$$\left(2m_n^2\right)^{-1}\exp\left(\frac{n\Delta_n}{(2+\varepsilon)v_n} - nc_1\right) + \exp\left\{\frac{n\Delta_n}{v_n}\left(\frac{1}{2+\varepsilon} - \frac{1}{2 + \frac{2}{3}(2m_n+1)v_n^{-1}\sqrt{\Delta_n}}\right)\right\}$$
$$\leq 1.$$

Since $\Delta_n \to 0$ and $v_n \geq var(b_1(U)b_1(V)) > 0$, it immediately follows that

$$\lim_{n\to\infty}\left(2m_n^2\right)^{-1}\exp\left(\frac{n\Delta_n}{(2+\varepsilon)v_n} - nc_1\right) = 0.$$

Using $\lim_{n\to\infty} m_n v_n^{-1}\sqrt{\Delta_n} = 0$, we get

$$\lim_{n\to\infty}\left(\frac{1}{2+\varepsilon} - \frac{1}{2 + \frac{2}{3}(2m_n+1)v_n^{-1}\sqrt{\Delta_n}}\right) = \frac{1}{2+\varepsilon} - \frac{1}{2} = -\frac{\varepsilon}{2(2+\varepsilon)},$$

which in combination with $n\Delta_n v_n^{-1} \geq n\Delta_n \|f\|_2^{-1} w_n^{-1}$ (see also (3.4)) and $\lim_{n\to\infty} n\Delta_n w_n^{-1} = \infty$ (see also Lemma 3.1 and (4.4)) gives

$$\lim_{n\to\infty} \exp\left\{\frac{n\Delta_n}{v_n}\left(\frac{1}{2+\varepsilon} - \frac{1}{2+\frac{2}{3}(2m_n+1)\,v_n^{-1}\sqrt{\Delta_n}}\right)\right\} = 0$$

and hence (4.10) holds for $n \geq n(\varepsilon)$. This completes the proof of the first inequality in (4.5) for $L \geq 1$. In case $L = 0$, the only difference is that the term $P\left(\widehat{c}^2_{R_L S_L} < \Delta_n\right)$ is not there; the remaining term $P\left(\widehat{c}^2_{R_{L+1} S_{L+1}} \geq \Delta_n\right)$ is covered by the proof for $L \geq 1$. The second inequality in (4.5) directly follows from $v_n \leq \|f\|_2 w_n$, cf. (3.4).

By (4.6) we have

$$\lim_{n\to\infty} \frac{n\Delta_n}{(\log m_n)^2} = \infty$$

and thus, using Lemma 3.1,

$$\lim_{n\to\infty}\left\{\log 2 + 2\log m_n - \frac{n\Delta_n}{(2+\varepsilon)\,\|f\|_2\, w_n}\right\} =$$
$$\lim_{n\to\infty} \log m_n \left\{\frac{\log 2}{\log m_n} + 2 - \frac{n\Delta_n}{(2+\varepsilon)\,\|f\|_2\, w_n \log m_n}\right\} = -\infty$$

and hence

$$\lim_{n\to\infty} 2m_n^2 \exp\left(-\frac{n\Delta_n}{(2+\varepsilon)\,\|f\|_2\, w_n}\right) = 0,$$

implying

$$\lim_{n\to\infty} P(S = L) = 1.$$

Next assume moreover that $n\Delta_n \geq \log n(\log m_n)^2$. In view of (4.5) to prove (4.7) it remains to show that for every $c > 0$

$$2m_n^2 \exp\left(-\frac{n\Delta_n}{(2+\varepsilon)\,\|f\|_2\, w_n}\right) = O(n^{-c})$$

as $n \to \infty$. Noting that $w_n = O(\log m_n)$, cf. Lemma 3.1, it follows that

$$\frac{n\Delta_n}{(2+\varepsilon)\,\|f\|_2\, w_n} \geq \frac{\log n(\log m_n)^2}{c_2 \log m_n} = c_2^{-1} \log n(\log m_n)$$

for some $c_2 > 0$. Therefore, for every $c > 0$,

$$2m_n^2 \exp\left(-\frac{n\Delta_n}{(2+\varepsilon)\,\|f\|_2\, w_n}\right) \leq \exp\left(-c_2^{-1}\log n(\log m_n) + \log 2 + 2\log m_n\right) = O(n^{-c})$$

as $n \to \infty$.

If $f_0 \in L_{2+\delta}$ for some $\delta > 0$ and hence also $f \in L_{2+\delta}$, then (3.5) together with Lemma 3.1 give $v_n = O(1)$ as $n \to \infty$ and hence (4.4) yields

$$\lim_{n\to\infty} 2m_n^2 \exp\left(-\frac{n\Delta_n}{(2+\varepsilon)\,\|f\|_2\, v_n}\right) = 0,$$

which in combination with the first inequality in (4.5) results in $\lim_{n\to\infty} P(S = L) = 1$.

Assuming moreover that $n\Delta_n \geq \log n(\log m_n)$ now gives, cf. (4.5), for some $c_3 > 0$,

$$P(S \neq L) \leq 2m_n^2 \exp\left(-\frac{n\Delta_n}{(2+\varepsilon)\,v_n}\right) \leq 2m_n^2 \exp\left(-\frac{\log n(\log m_n)}{c_3}\right)$$
$$= \exp\left(-c_3^{-1}\log n(\log m_n) + \log 2 + 2\log m_n\right) = O(n^{-c}) \text{ as } n \to \infty$$

and hence, for every $c > 0$,

$$P(S \neq L) = O(n^{-c}) \text{ as } n \to \infty,$$

thus completing the proof. ■

Conditions (4.4) are e.g. fulfilled if we take

$$\lim_{n\to\infty} m_n = \infty, \lim_{n\to\infty} \frac{m_n}{\sqrt{n}} \log n = 0 \text{ and } \Delta_n = n^{-1} (\log n) (\log m_n), \tag{4.11}$$

while conditions (4.6) are satisfied, when replacing $\lim_{n\to\infty} n^{-1/2}(\log n)m_n = 0$ in (4.11) by $\lim_{n\to\infty}(\log n)^{-1} \log m_n = 0$. In both cases we may take e.g. $m_n = O(\log n)$ as $n \to \infty$.

Remark 4.1 Obviously, the uniform copula density f_0^{un} satisfies $f_0 \in L_{2+\delta}$ for every $\delta > 0$. For the Gaussian copula density $f_0^G(u, v; \rho)$ we have $f_0^G \in L_{2+\delta}$ for every $0 < \delta \leq (1 - |\rho|)/|\rho|$ and hence for every $-1 < \rho < 1$ we obtain $f_0^G \in L_{2+\delta}$ for some $\delta > 0$.

4.2 Selecting the right Fourier coefficients

In this subsection we will show that the sets $\{(R_1, S_1), ..., (R_L, S_L)\}$ and $\{(r_1^*, s_1^*), ..., (r_L^*, s_L^*)\}$ coincide with probability tending to one. Let

$$A = \left\{ \min_{1\leq j\leq L} \left|\widehat{c}_{r_j^* s_j^*}\right| > \max_{L+1\leq j\leq m_n^2} \left|\widehat{c}_{r_j^* s_j^*}\right| \right\}. \tag{4.12}$$

On the set A the L largest $|\widehat{c}_{rs}|$ among $|\widehat{c}_{rs}|, 1 \leq r, s \leq m_n$ are $\left\{ \left|\widehat{c}_{r_1^* s_1^*}\right|, ..., \left|\widehat{c}_{r_L^* s_L^*}\right| \right\}$ and hence on the set A we have $\{(R_1, S_1), ..., (R_L, S_L)\} = \{(r_1^*, s_1^*), ..., (r_L^*, s_L^*)\}$. The following lemma gives an upper bound for $P(\overline{A})$.

Theorem 4.4 *Let f be given by (4.1) and L by (4.2). Then, for $L \geq 1$ and $m_n = o(n/\log n)$ as $n \to \infty$,*

$$\begin{aligned} P\left(\min_{1\leq j\leq L} \left|\widehat{c}_{r_j^* s_j^*}\right| \leq \max_{L+1\leq j\leq m_n^2} \left|\widehat{c}_{r_j^* s_j^*}\right| \right) &\leq 2m_n^2 \exp\left\{ - \frac{n c_{r_L^* s_L^*}^2}{8v_n + \frac{4}{3}\left|c_{r_L^* s_L^*}\right| (2m_n + 1)} \right\} \\ &= O\left(\exp\left\{ - \frac{n\left|c_{r_L^* s_L^*}\right|}{4m_n} \right\} \right), \end{aligned} \tag{4.13}$$

where $v_n = \max_{1\leq r,s\leq m_n} var(b_r(U)b_s(V))$. Hence, for every $c > 0$,

$$P(\{(R_1, S_1), ..., (R_L, S_L)\} = \{(r_1^*, s_1^*), ..., (r_L^*, s_L^*)\}) \geq 1 - O(n^{-c})$$

as $n \to \infty$.

Proof. Let $0 < \varepsilon < \left|c_{r_L^* s_L^*}\right|$. Noting that $E\widehat{c}_{r_j^* s_j^*} = c_{r_j^* s_j^*}$ for $1 \leq j \leq m_n^2$ and in particular, $E\widehat{c}_{r_j^* s_j^*} = 0$ for $L + 1 \leq j \leq m_n^2$, elementary inequalities yield

$$\begin{aligned} &P\left(\min_{1\leq j\leq L} \left|\widehat{c}_{r_j^* s_j^*}\right| \leq \max_{L+1\leq j\leq m_n^2} \left|\widehat{c}_{r_j^* s_j^*}\right| \right) \\ &\leq P\left(\min_{1\leq j\leq L} \left|\widehat{c}_{r_j^* s_j^*}\right| \leq \varepsilon \right) + P\left(\max_{L+1\leq j\leq m_n^2} \left|\widehat{c}_{r_j^* s_j^*}\right| \geq \varepsilon \right) \end{aligned} \tag{4.14}$$

$$
\begin{aligned}
&\le \sum_{j=1}^{L} P\left(\left|\widehat{c}_{r_j^* s_j^*}\right| \le \varepsilon\right) + \sum_{j=L+1}^{m_n^2} P\left(\left|\widehat{c}_{r_j^* s_j^*}\right| \ge \varepsilon\right) \\
&\le \sum_{j=1}^{L} P\left(\left|\widehat{c}_{r_j^* s_j^*} - E\widehat{c}_{r_j^* s_j^*}\right| \ge \left|c_{r_j^* s_j^*}\right| - \varepsilon\right) + \sum_{j=L+1}^{m_n^2} P\left(\left|\widehat{c}_{r_j^* s_j^*} - E\widehat{c}_{r_j^* s_j^*}\right| \ge \varepsilon\right) \\
&\le \sum_{j=1}^{L} P\left(\left|\widehat{c}_{r_j^* s_j^*} - E\widehat{c}_{r_j^* s_j^*}\right| \ge \left|c_{r_L^* s_L^*}\right| - \varepsilon\right) + \sum_{j=L+1}^{m_n^2} P\left(\left|\widehat{c}_{r_j^* s_j^*} - E\widehat{c}_{r_j^* s_j^*}\right| \ge \varepsilon\right).
\end{aligned}
$$

Take $\varepsilon = \frac{1}{2}\left|c_{r_L^* s_L^*}\right|$. In view of (4.8), cf. also(4.9), Bernstein's inequality gives

$$
\begin{aligned}
&P\left(\left|\widehat{c}_{r_j^* s_j^*} - E\widehat{c}_{r_j^* s_j^*}\right| \ge \frac{\left|c_{r_L^* s_L^*}\right|}{2}\right) \\
&\le 2\exp\left\{-\frac{n c_{r_L^* s_L^*}^2}{8v_n + \frac{4}{3}\left|c_{r_L^* s_L^*}\right|(2m_n+1)}\right\}.
\end{aligned} \tag{4.15}
$$

Combination of (4.14) and (4.15) leads to

$$
\begin{aligned}
&P\left(\min_{1\le j\le L}\left|\widehat{c}_{r_j^* s_j^*}\right| \le \max_{L+1\le j\le m_n^2}\left|\widehat{c}_{r_j^* s_j^*}\right|\right) \\
&\le \sum_{j=1}^{L} P\left(\left|\widehat{c}_{r_j^* s_j^*} - E\widehat{c}_{r_j^* s_j^*}\right| \ge \frac{\left|c_{r_L^* s_L^*}\right|}{2}\right) + \sum_{j=L+1}^{m_n^2} P\left(\left|\widehat{c}_{r_j^* s_j^*} - E\widehat{c}_{r_j^* s_j^*}\right| \ge \frac{\left|c_{r_L^* s_L^*}\right|}{2}\right) \\
&\le 2m_n^2 \exp\left\{-\frac{n c_{r_L^* s_L^*}^2}{8v_n + \frac{4}{3}\left|c_{r_L^* s_L^*}\right|(2m_n+1)}\right\},
\end{aligned}
$$

thus proving the first inequality of (4.13).

Noting that $v_n = O(w_n) = O(\log m_n)$, cf. (3.4) and Lemma 3.1, it follows that for sufficiently large n,

$$
8v_n + \frac{4}{3}\left|c_{r_L^* s_L^*}\right|(2m_n+1) \le 3\left|c_{r_L^* s_L^*}\right| m_n
$$

and hence, for sufficiently large n,

$$
\begin{aligned}
&2m_n^2 \exp\left\{-\frac{n c_{r_L^* s_L^*}^2}{8v_n + \frac{4}{3}\left|c_{r_L^* s_L^*}\right|(2m_n+1)}\right\} \\
&\le \exp\left\{-\frac{n c_{r_L^* s_L^*}^2}{3\left|c_{r_L^* s_L^*}\right| m_n} + \log(2m_n^2)\right\} \le \exp\left\{-\frac{n\left|c_{r_L^* s_L^*}\right|}{4m_n}\right\},
\end{aligned}
$$

where we have used that $m_n = o(n/\log n)$ as $n \to \infty$ and hence $n^{-1} m_n \log m_n \to 0$ as $n \to \infty$. Using again $m_n = o(n/\log n)$ as $n \to \infty$, it follows immediately that for every $c > 0$,

$$
\exp\left\{-\frac{n\left|c_{r_L^* s_L^*}\right|}{4m_n}\right\} = O(n^{-c})
$$

as $n \to \infty$, thus completing the proof. ■

4.3 Model error, stochastic error

In the general case of a parametric start we can distinguish three type of errors: (a) errors due to estimation of the parameter τ in the part concerning the parametric start; obviously, when starting with a given density f_0, as for instance the uniform density on the unit square f_0^{un}, such an error is not present; (b) the model error due to an error in selecting the model; (c) the stochastic error due to estimation of the Fourier coefficients within the selected model. Since $m_n \to \infty$ and K is fixed, we have for sufficiently large n that $m_n > K$. Therefore, without essential loss of generality we assume $m_n > K$. We use the following notation

$$\begin{aligned} g(u,v) &= f(u,v) - f_0(u,v;\tau) = \sum_{r=1}^{m_n}\sum_{s=1}^{m_n} c_{rs} b_r(u) b_s(v), \\ g_S(u,v) &= \sum_{j=1}^{S} c_{R_j S_j} b_{R_j}(u) b_{S_j}(v), \\ \widehat{g}_S(u,v;\tau) &= \sum_{j=1}^{S} \widehat{c}_{R_j S_j}(\tau) b_{R_j}(u) b_{S_j}(v). \end{aligned}$$

With this notation we can write

$$\begin{aligned} \widehat{f}(u,v) - f(u,v) = &\{f_0(u,v;\widehat{\tau}) - f_0(u,v;\tau) + \widehat{g}_S(u,v;\widehat{\tau}) - \widehat{g}_S(u,v;\tau)\} \\ &+ \{\widehat{g}_S(u,v;\tau) - g_S(u,v)\} + \{g_S(u,v) - g(u,v)\}. \end{aligned}$$

Here, we clearly see the various type of error terms. The first one, $f_0(u,v;\widehat{\tau}) - f_0(u,v;\tau) + \widehat{g}_S(u,v;\widehat{\tau}) - \widehat{g}_S(u,v;\tau)$, concerns the estimation of τ, that is estimation in the basic parametric model (if present). The second term, $\widehat{g}_S(u,v;\tau) - g_S(u,v)$ gives the error due to estimation within the selected parametric model. The third term, $g_S(u,v) - g(u,v)$ gives the model error due to restriction to dimension S in the orthonormal system. As stated before we will mainly concentrate on the second and third term.

As to the first term we have

$$\|f_0(\widehat{\tau}) - f_0(\tau) + \widehat{g}_S(\widehat{\tau}) - \widehat{g}_S(\tau)\|_2 \le \|f_0(\widehat{\tau}) - f_0(\tau)\|_2 + \|\widehat{g}_S(\widehat{\tau}) - \widehat{g}_S(\tau)\|_2$$

and

$$\begin{aligned} \|\widehat{g}_S(\widehat{\tau}) - \widehat{g}_S(\tau)\|_2^2 &= \left\| \sum_{j=1}^{S} \{\widehat{c}_{R_j S_j}(\widehat{\tau}) - \widehat{c}_{R_j S_j}(\tau)\} b_{R_j} b_{S_j} \right\|_2^2 \\ &= \sum_{j=1}^{S} \{\widehat{c}_{R_j S_j}(\widehat{\tau}) - \widehat{c}_{R_j S_j}(\tau)\}^2 \\ &= \sum_{j=1}^{S} \left[\iint b_{R_j}(u) b_{S_j}(v) \{f_0(u,v;\widehat{\tau}) - f_0(u,v;\tau)\} du dv \right]^2 \\ &\le \sum_{j=1}^{S} \iint \{b_{R_j}(u) b_{S_j}(v)\}^2 du dv \iint \{f_0(u,v;\widehat{\tau}) - f_0(u,v;\tau)\}^2 du dv \\ &= S \, \|f_0(\widehat{\tau}) - f_0(\tau)\|_2^2 , \end{aligned}$$

implying

$$\|f_0(\widehat{\tau}) - f_0(\tau) + \widehat{g}_S(\widehat{\tau}) - \widehat{g}_S(\tau)\|_2 \le (1 + \sqrt{S}) \, \|f_0(\widehat{\tau}) - f_0(\tau)\|_2 .$$

Since $S = L$ with probability tending (often at a very fast rate) to 1 (see Theorem 4.1), the order of the error due to the first term is indeed $\|f_0(\widehat{\tau}) - f_0(\tau)\|_2$, the estimation error in the basic parametric model (if present).

For the second and third term we obtain the following result.

Theorem 4.5 *We have*

$$\|\widehat{g}_S(\tau) - g\|_2^2 = \|\widehat{g}_S(\tau) - g_S\|_2^2 + \|g_S - g\|_2^2 \tag{4.16}$$
$$= \sum_{j=1}^{S} (\widehat{c}_{R_j S_j}(\tau) - c_{R_j S_j})^2 + \sum_{j=S+1}^{m_n^2} c_{R_j S_j}^2.$$

If

$$\lim_{n\to\infty} \Delta_n = 0, \lim_{n\to\infty} m_n = \infty, \lim_{n\to\infty} \frac{m_n \log n}{n} = 0, \lim_{n\to\infty} \frac{m_n\sqrt{\Delta_n}}{v_n} = 0, n\Delta_n \geq \log n (\log m_n)^2, \tag{4.17}$$

we get

$$E\,\|\widehat{g}_S(\tau) - g_S\|_2^2 = O(n^{-1}) \text{ as } n \to \infty, \tag{4.18}$$

and for every $c > 0$,

$$E\,\|g_S - g\|_2^2 = O(n^{-c}) \text{ as } n \to \infty. \tag{4.19}$$

If $f_0 \in L_{2+\delta}$ for some $\delta > 0$, then replacing $n\Delta_n \geq \log n(\log m_n)^2$ by $n\Delta_n \geq \log n(\log m_n)$ in (4.17) suffices for getting (4.18) and (4.19). So, if $f_0 \in L_{2+\delta}$ for some $\delta > 0$ and

$$\lim_{n\to\infty} \Delta_n = 0, \lim_{n\to\infty} m_n = \infty, \lim_{n\to\infty} \frac{m_n \log n}{n} = 0, \lim_{n\to\infty} \frac{m_n\sqrt{\Delta_n}}{v_n} = 0, n\Delta_n \geq \log n(\log m_n),$$

then

$$E\,\|\widehat{g}_S(\tau) - g\|_2^2 = O(n^{-1}) \text{ as } n \to \infty.$$

Proof. Orthonormality of the system $b_r(u)b_s(v)$ gives (4.16). By (4.8) we have for all $1 \leq r, s \leq m_n$,

$$|\widehat{c}_{rs} - c_{rs}| = \left| \frac{1}{n}\sum_{i=1}^{n} b_r(U_i)b_s(V_i) - \int\int b_r(u)b_s(v)dF(u,v) \right| \leq 2(2m_n + 1).$$

Write 1_A for the indicator function of the set A. Take A as in (4.12). Then,

$$E\left\{\sum_{j=1}^{S} (\widehat{c}_{R_j S_j}(\tau) - c_{R_j S_j})^2\right\}$$
$$= E\left\{\sum_{j=1}^{S} (\widehat{c}_{R_j S_j}(\tau) - c_{R_j S_j})^2 1_{A\cap\{S=L\}}\right\} + E\left\{\sum_{j=1}^{S} (\widehat{c}_{R_j S_j}(\tau) - c_{R_j S_j})^2 1_{\overline{A}\cup\{S\neq L\}}\right\}$$
$$\leq E\left\{\sum_{j=1}^{L} (\widehat{c}_{r_j^* s_j^*}(\tau) - c_{r_j^* s_j^*})^2 1_{A\cap\{S=L\}}\right\} + E\left\{\sum_{j=1}^{m_n} (\widehat{c}_{R_j S_j}(\tau) - c_{R_j S_j})^2 1_{\overline{A}\cup\{S\neq L\}}\right\}$$
$$\leq E\left\{\sum_{j=1}^{L} (\widehat{c}_{r_j^* s_j^*}(\tau) - c_{r_j^* s_j^*})^2\right\} + E\left\{4(2m_n+1)^2 m_n 1_{\overline{A}\cup\{S\neq L\}}\right\}$$
$$\leq \sum_{j=1}^{L} var\left(\widehat{c}_{r_j^* s_j^*}(\tau)\right) + 4(2m_n+1)^2 m_n\{P(\overline{A}) + P(S \neq L)\}.$$

Using $var(\widehat{c}_{rs}(\tau)) = n^{-1}var(b_r(U)b_s(V))$, we get, since L is fixed,

$$\sum_{j=1}^{L} var\left(\widehat{c}_{r_j^* s_j^*}(\tau)\right) = O(n^{-1}) \text{as } n \to \infty.$$

Assume that (4.17) holds. Using $\lim_{n\to\infty} n^{-1} m_n \log n = 0$, it follows from Theorem 4.4 and (4.7) that for every $c > 0$,

$$4(2m_n+1)^2 m_n \{P(\overline{A}) + P(S \neq L)\} = O(n^{-c}) \tag{4.20}$$

as $n \to \infty$, which completes the proof of (4.18).

Further we obtain

$$E\,\|g_S - g\|_2^2 = E\left(\sum_{j=S+1}^{m_n^2} c_{R_j S_j}^2\right) = E\left(\sum_{j=S+1}^{m_n^2} c_{R_j S_j}^2 1_{A\cap\{S=L\}}\right) + E\left(\sum_{j=S+1}^{m_n^2} c_{R_j S_j}^2 1_{\overline{A}\cup\{S\neq L\}}\right).$$

On the set A we have $\{(R_1,S_1),...,(R_L,S_L)\} = \{(r_1^*,s_1^*),...,(r_L^*,s_L^*)\}$ and hence

$$E\left(\sum_{j=S+1}^{m_n^2} c_{R_j S_j}^2 1_{A\cap\{S=L\}}\right) = E\left(\sum_{j=L+1}^{m_n^2} c_{r_j^* s_j^*}^2 1_{A\cap\{S=L\}}\right) = 0.$$

Moreover,

$$\begin{aligned} E\left(\sum_{j=S+1}^{m_n^2} c_{R_j S_j}^2 1_{\overline{A}\cup\{S\neq L\}}\right) &\le E\left(\sum_{j=1}^{m_n^2} c_{r_j^* s_j^*}^2 1_{\overline{A}\cup\{S\neq L\}}\right) = \|f-f_0\|_2^2\, E\left(1_{\overline{A}\cup\{S\neq L\}}\right) \\ &\le \|f-f_0\|_2^2\,\{P(\overline{A}) + P(S\neq L)\}, \end{aligned}$$

and thus, cf (4.20), (4.19) easily follows. If $f_0 \in L_{2+\delta}$ for some $\delta > 0$, the $(\log m_n)^2$-term in (4.17) may be replaced by a $\log m_n$-term in view of Theorem 4.1. Combination of (4.16), (4.18) and (4.19) gives the last statement of the theorem. ■

Remark 4.2 It is seen from Theorems 3.2 and 4.5 that the selection rule S reduces the expected quadratic error with a factor m_n^2 and brings it back to the classical $O(n^{-1})$.

Remark 4.3 From a pure theoretical point of view, the "optimal" choice of Δ_n and m_n for densities f of the form (4.1) is simply: take Δ_n as large as possible and m_n as small as possible. In fact, if we would know K and $c_{r_L^* s_L^*}$, taking Δ_n as a constant smaller than $\left|c_{r_L^* s_L^*}\right|$ and $m_n = K$ gives that both $P(\overline{A})$ and $P(S \neq L)$ are exponentially small. However, we do not know K and $c_{r_L^* s_L^*}$ and therefore the "optimal" choice is to take $\Delta_n \to 0$ and $m_n \to \infty$ as slow as possible. On the other hand, the error given by (4.18) is $O(n^{-1})$ and from this perspective, other choices for Δ_n and m_n are already good enough. Therefore, it is interesting to cover in the theory also more classical choices of Δ_n coming from model selection theory with some adaptation for the fact that we go through the dimensions in a data driven way, taking the largest Fourier coefficients first. This leads to the choice $\Delta_n = n^{-1}(\log n)(\log m_n)$. For m_n we may take $m_n = \log n$, thus tending to infinity not too fast. With this choice we get for densities f of the form (4.1) with $f_0 \in L_{2+\delta}$ for some $\delta > 0$ that $P(\overline{A})$ and $P(S \neq L)$ are smaller than n^{-c} for every $c > 0$, when n is sufficiently large.

5 Application

In this section the method is illustrated on a real life example concerning 1500 U.S. insurance claim data. Each pair of variables consists of an indemnity payment ($LOSS$) and an allocated loss adjustment expense ($ALAE$), see e.g. Frees and Valdez (1998), Klugman and Parsa (1999) for more explanation about these data. A picture of $log(LOSS)$ against $log(ALAE)$ is given in Figure 1.

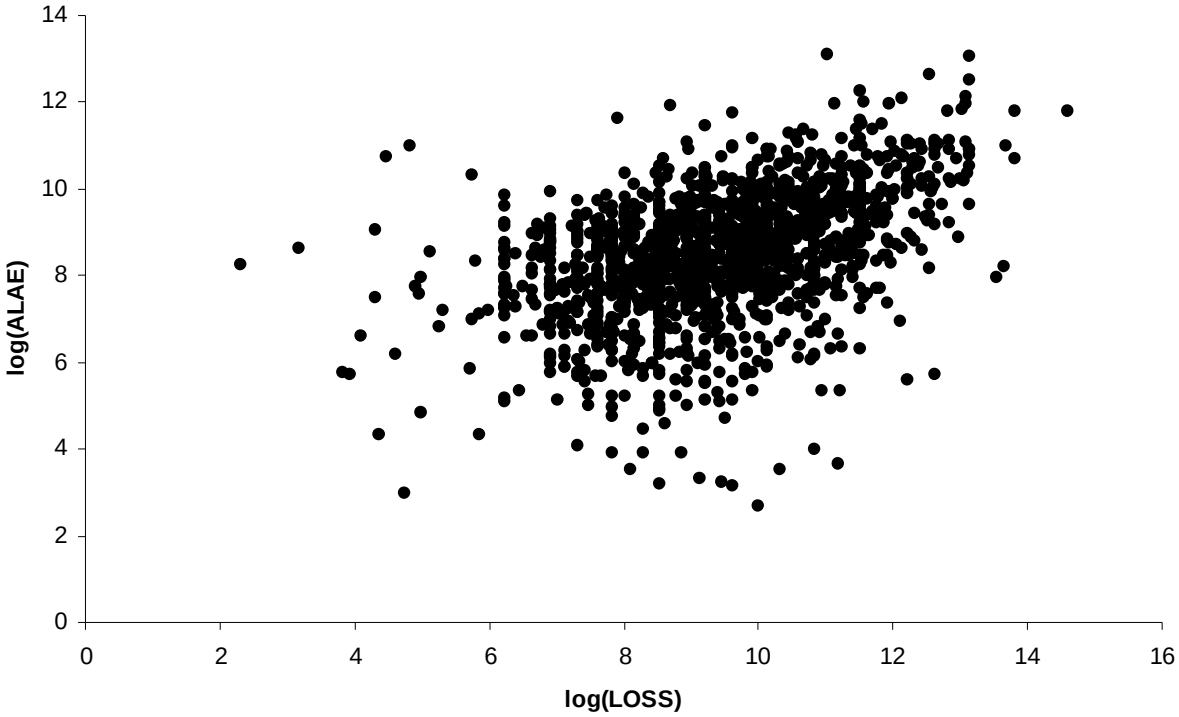

Figure 1. $\log(LOSS)$ *against* $\log(ALAE)$.

To obtain (approximately) uniform marginals we apply on both variables their fitted marginal distribution function. As mentioned in the Introduction the topic of the present paper is estimating the copula and not fitting the marginals. Therefore we simply apply the (nonparametric) empirical distribution functions on the marginals. This gives our "observations" U_i and V_i on which we want to illustrate the new method of estimating a copula.

When considering the Gaussian copula we have to estimate the correlation coefficient ρ, see (2.1). Because $\rho = \rho(\Phi^{-1}(U), \Phi^{-1}(V))$, the natural estimator of ρ is the sample correlation coefficient based on $(\Phi^{-1}(U_i), \Phi^{-1}(V_i))$. To avoid problems with $U_i = 1$ or $V_i = 1$, we take as empirical distribution function for X: $F_n^X(x) = (n+1)^{-1} \sum_{i=1}^n 1(X_i \leq x)$ and similarly for Y. Hence, we get as basic observations

$$U_i = F_{1500}^X(X_i), V_i = F_{1500}^Y(Y_i), i = 1, ..., 1500.$$

A picture of the U_i and V_i is given in Figure 2.

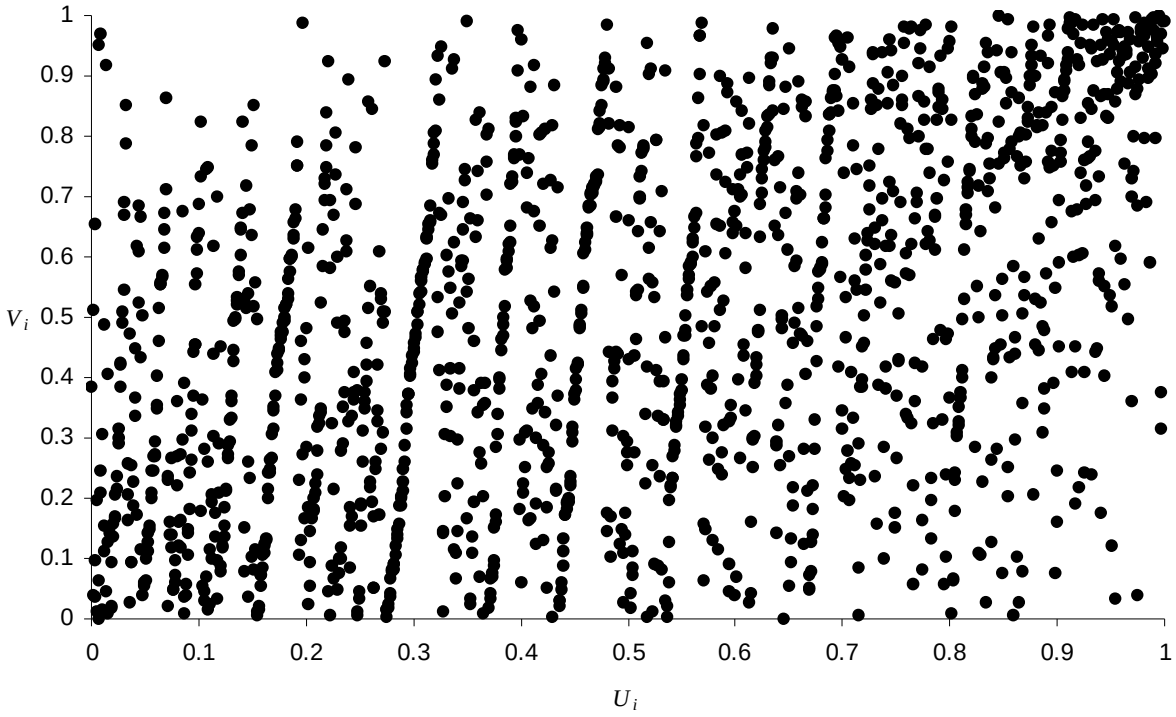

Figure 2. LOSS against ALAE with (estimated) uniform marginals.

The estimated copula is given by, cf. (3.9),

$$\widehat{f}(u,v) = f_0(u,v;\widehat{\tau}) + \sum_{j=1}^{S} \widehat{c}_{R_jS_j}(\widehat{\tau}) b_{R_j}(u) b_{S_j}(v).$$

Firstly, we consider the uniform start with $f_0(u,v;\widehat{\tau}) = 1$ for all u, v. In that case there is no additional parameter τ and $\widehat{c}_{rs}(\widehat{\tau}) = \widehat{c}_{rs} = n^{-1}\sum_{i=1}^{n} b_r(U_i)b_s(V_i)$. We take $m_{1500} = 10$ and, cf. (3.8), $\Delta_{1500} = 1500^{-1}(\log 1500)(\log 10) = 0.0112$. To get the number of terms S in our approximation and the coefficients $\widehat{c}_{R_jS_j}$, we have to calculate $\widehat{c}_{rs}$ for $r, s = 1, ..., 10$, order their squares and take those for which $\widehat{c}_{rs}^2 \geq \Delta_{1500}$, cf. (3.7), or, equivalently, for which $|\widehat{c}_{rs}| \geq \sqrt{\Delta_{1500}} = 0.1060$. Direct calculation gives that $|\widehat{c}_{rs}| \geq \sqrt{\Delta_{1500}}$ for $(r,s) = (1,1), (2,2), (1,2)$ and $(2,3)$ with corresponding values:

$$\widehat{c}_{11} = 0.4624, \widehat{c}_{22} = 0.2185, \widehat{c}_{12} = 0.1250 \text{ and } \widehat{c}_{23} = 0.1215.$$

Hence, we get $S = 5, (R_1, S_1) = (1,1), (R_2, S_2) = (2,2), (R_3, S_3) = (1,2), (R_4, S_4) = (2,3)$ and $(R_5, S_5) = (4,2)$. The resulting approximation, denoted by $\widehat{f}^{un}$ because of the uniform start, therefore is

$$\widehat{f}^{un}(u,v) = 1 + 0.4624 b_1(u)b_1(v) + 0.2185 b_2(u)b_2(v) + 0.1250 b_1(u)b_2(v) + 0.1215 b_2(u)b_3(v). \quad (5.1)$$

A picture of the estimated copula $\widehat{f}$ is presented in Figure 3.

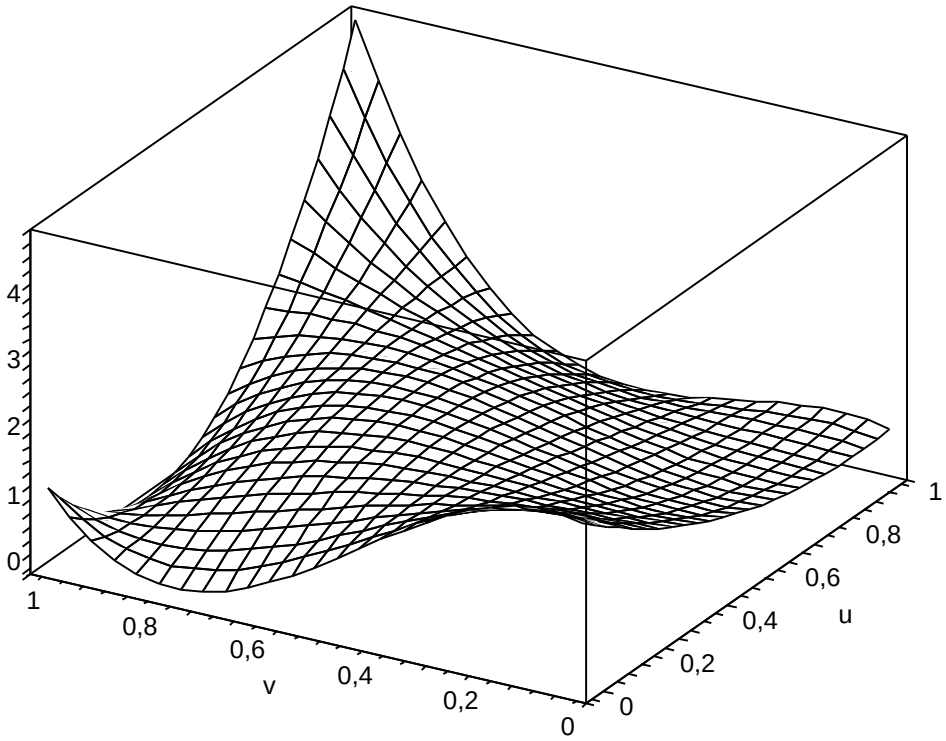

Figure 3. Estimated copula with uniform start.

For the start with the Gaussian copula we estimate ρ by the sample correlation coefficient

$$\widehat{\rho} = \frac{\sum_{i=1}^{1500} \left\{\Phi^{-1}(U_i) - \overline{\Phi^{-1}(U)}\right\}\left\{\Phi^{-1}(V_i) - \overline{\Phi^{-1}(V)}\right\}}{\sqrt{\sum_{i=1}^{1500} \left\{\Phi^{-1}(U_i) - \overline{\Phi^{-1}(U)}\right\}^2 \sum_{i=1}^{1500} \left\{\Phi^{-1}(V_i) - \overline{\Phi^{-1}(V)}\right\}^2}},$$

where

$$\overline{\Phi^{-1}(U)} = \frac{\sum_{i=1}^{1500} \Phi^{-1}(U_i)}{1500}.$$

This results in $\widehat{\rho} = 0.4756$. Now we have

$$\widehat{c}_{rs} = \widehat{c}_{rs}(\widehat{\rho}) = \frac{1}{1500}\sum_{i=1}^{1500} b_r(U_i)b_s(V_i) - \iint b_r(u)b_s(v)f_0(u,v;0.4756)dudv.$$

Direct calculation gives that $|\widehat{c}_{rs}| \geq \sqrt{\Delta_{1500}}$ for $(r,s) = (1,2)$ and $(2,3)$ with corresponding values:

$$\widehat{c}_{12} = 0.1250 \text{ and } \widehat{c}_{23} = 0.1215.$$

We see that the linear correlation, which was the highest Fourier coefficient when starting with the uniform, is already taken into account by the Gaussian copula, as may be expected because of the extra parameter ρ, which is involved. The coefficients $\widehat{c}_{12}$ and $\widehat{c}_{23}$ are the same as in the uniform case, because $\iint b_r(u)b_s(v)f_0(u,v;\rho)dudv = 0$ for $r+s$ odd, due to the fact that $b_r(u)b_s(v) = -b_r(1-u)b_s(1-v)$ if $r+s$ is odd and $f_0(u,v;\rho) = f_0(1-u,1-v;\rho)$, implying $\widehat{c}_{rs}(\widehat{\rho}) = 1500^{-1}\sum_{i=1}^{1500} b_r(U_i)b_s(V_i)$ for $r+s$ odd, as in case of a uniform start. The resulting approximation, denoted by $\widehat{f}^G$, because of the Gaussian start, therefore is

$$\widehat{f}^G(u,v) = f_0(u,v;0.4756) + 0.1250b_1(u)b_2(v) + 0.1215b_2(u)b_3(v).$$

A picture of the estimated copula $\widehat{f}^G$ is presented in Figure 4.

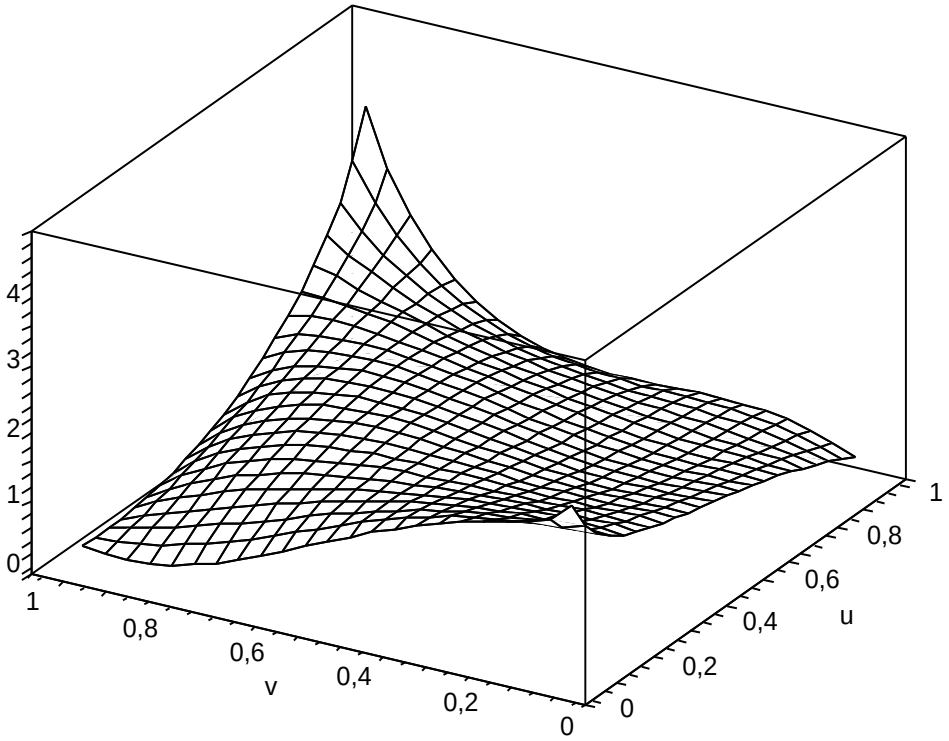

Figure 4. Estimated copula with Gaussian start.

To illustrate the performance of the estimated copulas we compare frequencies of the data with estimated probabilities for the same rectangles as used in Kallenberg (2008), that is symmetric ones with smaller and larger frequencies ($u = v = 0.25, 0.4$) and asymmetric ones ($u = 0.25, v = 0.5; u = 0.5, v = 0.25$) and also the corresponding upper tail rectangles.

Table 1 *Frequencies and approximations on various rectangles.*

rectangle	$freq$	$f_0^{un}/freq$	$f_0^G/freq$	$\widehat{f^{un}}/freq$	$\widehat{f^G}/freq$
$(0, 0.25) \times (0, 0.25)$	0.1087	0.575	1.076	1.027	0.991
$(0, 0.4) \times (0, 0.4)$	0.2240	0.714	1.048	1.065	1.031
$(0, 0.25) \times (0, 0.5)$	0.1800	0.694	1.036	1.079	1.060
$(0, 0.5) \times (0, 0.25)$	0.1807	0.692	1.032	0.989	0.970
$(0.75, 1) \times (0.75, 1)$	0.1333	0.469	0.877	0.976	0.947
$(0.6, 1) \times (0.6, 1)$	0.2420	0.661	0.970	1.018	0.987
$(0.75, 1) \times (0.5, 1)$	0.1840	0.679	1.014	1.010	0.991
$(0.5, 1) \times (0.75, 1)$	0.1980	0.631	0.942	1.017	0.999
mean abs. rel. diff.		36.0%	5.2%	3.1%	2.6%

It is clearly seen from Table 1 that the uniform approximation f_0^{un}, ignoring the dependence at all, is very bad. The classical Gaussian approximation f_0^G gives an enormous improvement. Although the uniform start is very bad, nevertheless the approximation with the uniform start $\widehat{f^{un}}$ gives very good results with often a substantial improvement compared to the classical Gaussian approximation. The approximation with the Gaussian start $\widehat{f^G}$ gives similar results as the approximation with the uniform start: again a substantial improvement w.r.t. the classical Gaussian approximation with a slightly further improvement w.r.t. the approximation with the uniform start. This is also seen in the last line of the table, where the mean of the absolute values of the relative differences is given. So, for instance 5.2% stands for the mean of $\left|f_0^G/freq - 1\right|$ over the 8 cases. There we see the enormous improvement from 36% to 5.2%, when taking the

classical Gaussian copula (obviously due to the high amount of linear correlation in the data) and the substantial further improvement reducing the mean with 50% from 5.2 % to 2.6% when applying the contamination approximation with a Gaussian start.

As an example of an extreme quantile, consider the 99%-quantile, given by the rectangle $(0,u)\times(0,u)$ such that $\int_0^u\int_0^u \widehat{f}^{un}=0.99$. This gives $u=0.9949$. So, the expected number of data points outside this rectangle equals $(1-0.99)\times 1500=15$. The actual number of data points (U_i,V_i) outside the rectangle $(0,0.9949)\times(0,0.9949)$ equals 13, which is very close to the expected one. If we consider in a similar way the 99%-quantile in the (X,Y)-plane with $X=\log(LOSS)$ and $Y=log(ALAE)$, that is we approximate the 99%-quantile $\widetilde{x}$ satisfying $F_{X,Y}(\widetilde{x},\widetilde{x})=0.99$, or, equivalently, $P(X\leq\widetilde{x},Y\leq\widetilde{x})=P(F^X_{1500}(X)\leq F^X_{1500}(\widetilde{x}),F^Y_{1500}(Y)\leq F^Y_{1500}(\widetilde{x}))=P(U\leq F^X_{1500}(\widetilde{x}),V\leq F^Y_{1500}(\widetilde{x}))=0.99$ by solving $\int_0^{\widetilde{F}^X_{1500}(x)}\int_0^{\widetilde{F}^Y_{1500}(x)}\widehat{f}^{un}=0.99$ (with $\widetilde{F}^X_{1500}$ and $\widetilde{F}^Y_{1500}$ the linearized version of F^X_{1500} and F^Y_{1500}, respectively), we get $x=13.1155$ (and $\widetilde{F}^X_{1500}(x)=0.9907,\widetilde{F}^Y_{1500}(x)=0.9992$). So 13.1155 is the estimated 99%-quantile based on the estimated density $\widehat{f}^{un}$. The actual number of data points (X_i,Y_i) outside the rectangle $(-\infty,13.1155)\times(-\infty,13.1155)$ equals 14. Hence, also in the (X,Y)-plane the actual number of data points is very close to the expected number of data points predicted by using $\widehat{f}^{un}$. Replacing $\widehat{f}^{un}$ by $\widehat{f}^{G}$ gives the same results. This shows that also the extreme quantiles are very well estimated by using $\widehat{f}^{un}$ or $\widehat{f}^{G}$.

We may conclude that the contamination family both with a uniform start and with a Gaussian start give very nice results approximating the data with high accuracy.

References

Biau, G. and Wegkamp, M. (2005). A note on minimum distance estimation of copula densities. *Statist. Probab. Lett.* **73** 105-114.

Cherubini, U., Luciano, E. and Vecchiato, W. (2004). *Copula Methods in Finance.* Wiley, Chichester.

Embrechts, P., Lindskog, F. and McNeil, A. (2003). Modelling Dependence with Copulas and Applications to Risk Management. In: *Handbook of Heavy Tailed Distributions in Finance* (S. T. Rachev, ed.) 329-384, Elsevier, Amsterdam.

Embrechts, P, McNeil, A.J. and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. In *Risk management: value at risk and beyond* (M. A. H. Dempster, ed.) 176-223, Cambridge Univ. Press, Cambridge.

Fermanian, J.-D. (2005). Goodness of fit tests for copulas. *J. Multivariate Anal.* **95** 119-152.

Frees, E.W., Valdez, E.A. (1998). Understanding relationships using copulas. *N. Am. Actuar. J.* **2** 1-25.

Genest, C. and Favre, A.-C. (2007). Everything you always wanted to know about copula modeling but were afraid to ask. Journal of Hydrologic Engineering, **12**, 347-368.

Genest, C. and Rémillard, B. (2006). Discussion of: "Copulas: tales and facts" by T. Mikosch. *Extremes* **9** 27-36.

Genest, C., Rémillard, B. and Beaudoin, D. (2008). Goodness-of-fit tests for copulas: A review and a power study. *Insurance Math. Econom.* **42**, In Press.

Joe, H.(1997). *Multivariate Models and Dependence Concepts* Monographs on Statistics and Applied Probability **73** Chapman & Hall, London.

Kallenberg, W.C.M. (2008), Modelling dependence. *Insurance Math. Econom.* **42** 127-146.

Kallenberg, W.C.M. and Ledwina, T. (1997). Data-driven smooth tests when the hypothesis is composite. *J. Amer. Statist. Assoc.* **92** 1094-1104.

Kallenberg, W.C.M. and Ledwina, T. (1999). Data-driven rank tests for independence. *J. Amer. Statist. Assoc.* **94** 285-301.

Klugman, S.A., Parsa, R. (1999) Fitting bivariate loss distributions with copulas. *Insurance Math. Econom.* **24** 139-148.

Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.* **89** 1000-1005.

Mikosch, T. (2006). Copulas: tales and facts *Extremes* **9** 3-20.

McNeil, A., Frey, R. and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, Princeton.

Nelsen, R. B. (1999). *An Introduction to Copulas.* Lecture Notes in Statistics **139** Springer-Verlag, New York.

Panchenko V. (2005). Goodness-of-fit test for copulas. *Phys. A* **355** 176-182.

Sansone, G. (1959). *Orthogonal Functions*. Interscience, New York.

Sklar, A., (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* **8** 229-231.

Sklar, A. (1996). Random variables, distribution functions, and copulas – a personal look backward and forward. In *Distributions with Fixed Marginals and Related Topics* (L. Rüschendorf, B. Schweizer and M. D. Taylor, eds).1-14, Lecture notes monograph series **28**, Institute of Mathematical Statistics, Hayward, CA.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.