Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 2836–2850
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Comparison of semiparametric and parametric methods for estimating copulas☆

Gunky Kim, Mervyn J. Silvapulle*, Paramsothy Silvapulle

Department of Econometrics and Business Statistics, Monash University, Caulfield East, Melbourne 3145, Australia

Received 22 December 2005; received in revised form 4 October 2006; accepted 4 October 2006
Available online 7 November 2006

Abstract

Copulas have attracted significant attention in the recent literature for modeling multivariate observations. An important feature of copulas is that they enable us to specify the univariate marginal distributions and their joint behavior separately. The copula parameter captures the intrinsic dependence between the marginal variables and it can be estimated by parametric or semiparametric methods. For practical applications, the so called inference function for margins (IFM) method has emerged as the preferred fully parametric method because it is close to maximum likelihood (ML) in approach and is easier to implement. The purpose of this paper is to compare the ML and IFM methods with a semiparametric (SP) method that treats the univariate marginal distributions as unknown functions. In this paper, we consider the SP method proposed by Genest et al. [1995. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 82(3), 543–552], which has attracted considerable interest in the literature. The results of an extensive simulation study reported here show that the ML/IFM methods are nonrobust against misspecification of the marginal distributions, and that the SP method performs better than the ML and IFM methods, overall. A data example on household expenditure is used to illustrate the application of various data analytic methods for applying the SP method, and to compare and contrast the ML, IFM and SP methods. The main conclusion is that, in terms of statistical computations and data analysis, the SP method is better than ML and IFM methods when the marginal distributions are unknown which is almost always the case in practice.

Keywords: Canonical maximum likelihood; Dependence; Inference function for margins; Portfolio management; Pseudo maximum likelihood

1. Introduction

There has been a growing interest in modeling multivariate observations using flexible functional forms for distribution functions and in estimating parameters that capture the dependence among different components. One of the main reasons for such interest is that the traditional approach based on multivariate normal distribution is not flexible enough because it is capable of representing only a very limited range of distributional shapes. Consequently, it has been widely recognized that there is a genuine need for other methods of modeling multivariate data. In response, methods based on *copula* have been the subject of extensive study in the recent literature.

☆ This research was partially supported by an Australian Research Council Discovery Project Grant and a Monash University Postgraduate Student Award.
* Corresponding author. Tel.: +61 3 99032237; fax: +61 3 99032007.
E-mail address: mervyn.silvapulle@buseco.monash.edu.au (M.J. Silvapulle).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.10.009

For a short introduction to copulas, see Genest and MacKay (1986). For book-length expositions that are oriented towards mathematical statistics, see Joe (1997) and Nelsen (2006). For expositions that are specifically oriented towards applications in economics, finance and risk management, see Cherubini et al. (2004), McNeil et al. (2005), Patton (2004, 2005a,b) and Granger et al. (2005). This list and the references therein provide a wealth of references to theory and applications.

In what follows, we shall restrict our discussion to bivariate observations only for simplicity. However, the main findings would be just as relevant in higher dimensions as well. Let (X_1, X_2) be a continuous bivariate random variable, $H(x_1, x_2)$ denote the *cdf* of (X_1, X_2), and let F_k and f_k denote the marginal *cdf* and *pdf*, respectively, of X_k, $(k = 1, 2)$. Then, a well-known result due to Sklar (1959) says that there is a unique function $C(u_1, u_2)$, termed the *copula*, such that

$$H(x_1, x_2) = C\{F_1(x_1), F_2(x_2)\}. \tag{1}$$

It turns out that the copula $C(\cdot, \cdot)$ is the joint distribution of (U_1, U_2) where, $U_k = F_k(X_k)$, $k = 1, 2$; clearly, U_1 and U_2 are uniformly distributed on (0, 1). Thus, any continuous bivariate distribution is uniquely defined by its marginal distributions and its copula. Conversely, given the univariate *cdf*s F_1 and F_2 and a joint *cdf* $C(u_1, u_2)$ on the unit square, it turns out that $C\{F_1(x_1), F_2(x_2)\}$ is the bivariate *cdf* with marginal distributions F_1 and F_2 and copula $C(u_1, u_2)$. Thus, the approach based on copulas, visualizes the joint distributions in terms of variables that have been transformed to have uniform (0,1) distributions. This helps to separate out the role of the marginal distributions from the intrinsic dependence between the components.

These results suggest the flexible approach of specifying a bivariate distribution by specifying the marginal distributions and the copula separately. For example, it is possible to specify: (a) a gamma distribution for X_1; (b) a t-distribution for X_2, and (c) a copula to capture the joint behavior of the two variables. These three stages can be performed in any order and independent of each other. The shapes of the marginal distributions of X_1 and X_2 do not play a role in the specification of the copula. For example, if $Y_k = h_k(X_k)$ where h_k is continuous and increasing $(i = 1, 2)$, then the copula of (X_1, X_2) is the same as that of (Y_1, Y_2). This illustrates the important point that the copula captures features of the joint distribution that are invariant under monotonic transformations of the marginal variables. Consequently, the units of measurement do not affect the copula. A main reason for the richness and flexibility of the copula based approach is that there are a large number of families of copulas documented and available for our choice. We shall write $C(u_1, u_2; \theta)$ for a family of copulas indexed by the parameter θ. Bivariate normal and bivariate t are of course special cases.

In this paper, we focus on estimating the parameter θ of the copula. In general, the maximum likelihood estimator (MLE) would be the preferred first option for estimating θ in view of its well-known optimality properties. However, the more flexible inference function for margins (IFM) method (see Joe, 1997, 2005) has emerged as preferable to ML because the conceptual basis for the two are very close, IFM method is easier than the ML method for computations, and the two methods are almost equally efficient in many cases. Since IFM is a fully parametric method, misspecification of the marginal distributions may have an effect on the performance of the estimator. In this paper, we study the robustness properties of ML and IFM estimators against misspecification of the marginal distributions and compare them with a semiparametric (SP) method that treats the marginal distributions as unknown functions.

The SP method that we chose to study is the one proposed by Genest et al. (1995). This is also known as *pseudo maximum likelihood* (PML) and as *canonical maximum likelihood* (see Cherubini et al., 2004). In what follows, we shall use the former term. This general approach has been studied in different contexts with modifications (for example, Shih and Louis, 1995; Wang and Ding, 2000; Tsukahara, 2005). A result in Genest et al. (1995) says that the PMLE is not as efficient as the MLE in general, and Genest and Werker (2002) obtained conditions under which the PMLE is asymptotically efficient.

The PML method estimates each marginal distribution nonparametrically by the empirical distribution function (*edf*), thus allowing the distribution of the marginals to be quite free and not restricted by parametric families. Once this is done, the interdependence between the margins is estimated using a parametric family of copulas. In contrast to parametric methods such as the maximum likelihood (ML) and IFM, an attractive feature of the PML method is that it does not require the user to specify functional forms for the marginal distributions.

Theoretical comparison of ML and IFM estimators with PMLE are difficult. Therefore, our comparisons are based on a well-designed simulation study. The results presented in this paper provide strong evidence to support the claim that PML is better than the well-known IFM and ML methods when the marginal distributions are unknown, which

is almost always the case in practice. This is the main contribution of this paper. Another attractive feature of PML method is that it is easy to code and also to implement in computer software.

Let us note that the objective of this paper is to highlight some of the important advantages of a SP approach compared to the well-known parametric ones in this area. To this end, we chose the SP estimator proposed by Genest et al. (1995). However, other methods such as the RAZ-estimator proposed by Tsukahara (2005), might have served just as well. In a simulation study, Tsukahara (2005) observed that the RAZ-estimator and the estimator of Genest et al. (1995) were the two best performing ones and none was uniformly the best. A comparison of the different SP estimators is outside the main objective of this paper, and therefore we restrict our study to one that has been well established in the literature.

The structure of the remaining sections is as follows. Section 2 provides an outline of the problem and several competing methods of inference. The simulation results are given in Section 3, an illustrative example is discussed in Section 4, Section 5 provides a discussion of the results, and Section 6 concludes the paper.

2. Specification and estimation of copulas

Let (X_1, X_2) denote a continuous bivariate random variable, and $F_k(x; \alpha_k)$ and $f_k(x; \alpha_k)$ be the *cdf* and *pdf*, respectively of X_k. The parameter α_k may be a vector, in which case we shall denote its transpose by α_k'. Let $U_k = F_k(X_k; \alpha_k)$, $C(u_1, u_2; \theta)$ denote the joint *cdf* of (U_1, U_2) (in other words the copula of (X_1, X_2)), $c(u_1, u_2; \theta)$ denote the *pdf* corresponding to $C(u_1, u_2; \theta)$, $\xi = (\alpha_1', \alpha_2', \theta')'$, and $H(x_1, x_2; \xi)$ and $h(x_1, x_2; \xi)$ denote the *cdf* and *pdf* of (X_1, X_2), respectively. For the most part, we shall consider the case when θ is a scalar, although an extension to the vector case would be straightforward. In this paper, we are interested in estimating θ using iid observations (X_{1i}, X_{2i}), $i = 1, \ldots, n$. Let us first mention briefly the methods of estimating θ that are considered here.

2.1. Maximum likelihood [ML]

In view of (1), the joint density function $h(x_1, x_2; \xi)$ of (X_1, X_2) can be expressed as follows:

$$h(x_1, x_2; \xi) = c\{F_1(x_1; \alpha_1), F_2(x_2; \alpha_2); \theta\} f_1(X_1; \alpha_1) f_2(X_2; \alpha_2). \tag{2}$$

Therefore, the loglikelihood function takes the form

$$L(\xi) = \sum_{i=1}^{n} \log[c\{F_1(X_{1i}, \alpha_1), F_2(X_{2i}; \alpha_2); \theta\} f_1(X_{1i}; \alpha_1) f_2(X_{2i}; \alpha_2)]. \tag{3}$$

Hence the *MLE* of ξ, which we denote by ξ^{**} is the global maximizer of $L(\xi)$. Then, we have that $\sqrt{n}(\xi^{**} - \xi_0)$ converges to a normal distribution with mean zero, where ξ_0 is the true value. Since we assume that the model is correctly specified and hence $L(\xi)$ is the correct loglikelihood, it follows that the *MLE* enjoys some optimality properties and hence is the preferred first option. If the model is not correctly specified so that $L(\xi)$ is not the correct loglikelihood, then the maximizer of $L(\xi)$ is not the MLE and hence it may lose its preferred status.

2.2. Inference function for margins [IFM]

In this method, the parameters are estimated in two stages. In the first stage, each marginal distribution is estimated separately. Thus, α_k is estimated using $X_{k1}, \ldots, X_{kn}$. Let this estimator be denoted by $\hat{\alpha}_k$ $(k = 1, 2)$. Then, in the second stage, θ is estimated by substituting $\hat{\alpha}_k$ for α_k in the loglikelihood and then maximizing the resulting function. Thus, the IFM estimate of θ is the maximizer of

$$\sum_{i=1}^{n} \log\left[c\{F_1(X_{1i}, \hat{\alpha}_1), F_2(X_{2i}; \hat{\alpha}_2); \theta\}\right]; \tag{4}$$

let us denote this estimator by $\hat{\theta}$. Under a reasonable set of regularity conditions, we have that $\sqrt{n}(\hat{\theta} - \theta_0)$ is asymptotically normal with mean zero; for example, see Joe (1997, Chapter 10).

2.3. Semiparametric method [SP]

The *ML* and *IFM* methods just mentioned are completely parametric because they require the model to be specified up to a finite number of unknown parameters. A possible shortcoming of these two methods of estimating θ is that they are likely to be inconsistent even if just one marginal distributions is misspecified. In this SP method, we allow the marginal distributions to have arbitrary and unknown functional forms. Estimation is carried out in two stages as in IFM, but the difference is that the marginal distributions are estimated nonparametrically by their sample empirical distributions. More specifically, let $\tilde{F}_k$ denote the *empirical cdf* of $X_{k1}, \ldots X_{kn}$, $(k = 1, 2)$. Then, θ is estimated by the maximizer of the *pseudo loglikelihood*,

$$\sum_{i=1}^{n} \log\left[c\{\tilde{F}_1(X_{1i}), \tilde{F}_2(X_{2i}); \theta\}\right]. \tag{5}$$

Let us denote the resulting (SP) estimator by $\tilde{\theta}$. In what follows, we shall refer to this as the *pseudo maximum likelihood estimator* (PMLE) of θ.

It has been shown that $\sqrt{n}(\tilde{\theta} - \theta_0)$ is asymptotically $N(0, v^2)$ for some v^2 (for example, see Genest et al., 1995; Tsukahara, 2005). Further, a consistent estimator of the joint *cdf* of (X_1, X_2) is $C(\tilde{F}_1(x_1), \tilde{F}_2(x_2); \tilde{\theta})$. These results hold whether or not we know the marginal distributions. A large sample 95% confidence interval for θ is $\tilde{\theta} \pm 1.96\tilde{v}$ where $\tilde{v}$ is a consistent estimator of v; one such estimator is given in Section 3 of Genest et al. (1995).

2.4. A bench-mark estimator [BM]

In order to evaluate the performance of the foregoing estimators, we introduce the following artificial benchmark (BM) estimator. Let F_1 and F_2 be as in the previous subsection. Let us suppose that these distribution functions are known, and let θ be estimated by the maximizer of the loglikelihood

$$\sum_{i=1}^{n} \log\left[c\{F_1(X_{1i}), F_2(X_{2i}); \theta\}\right]. \tag{6}$$

Let us denote the resulting estimator by θ^*. Note that the difference between (5) and (6) is that in (5) F_k is replaced by $\tilde{F}_k$. Although, the marginal distributions are unknown in practice, this hypothetical scenario, wherein F_k is assumed to be known, represents the ideal situation that can be used as a benchmark for comparative purposes. We would not expect ML/IF/SP estimators to perform better than θ^*. The difference between the efficiencies of θ^* and the estimators in the previous subsections, quantify the loss due to the marginal distributions being unknown.

3. Simulation study

A simulation study was carried out to compare the different estimators mentioned in the previous section for a range of copulas and marginal distributions. The two main objectives of the study are: (1) compare ML, IFM and PML estimators in terms of mean squared error (MSE) when the marginal distributions for ML and IFM may be misspecified, and (2) estimate the coverage rate of a large sample 95% confidence interval of the form $\tilde{\theta} \pm 1.96\tilde{v}$. In this simulation study, we used the $\tilde{v}$ in Section 3 of Genest et al. (1995) although we could have used any consistent estimator of v.

3.1. Design of the simulation study

The six copulas that we studied and the corresponding functional forms of $C(u, v; \theta)$ are given below:

(1) *Ali–Mikhail–Haq (AMH) Family of copulas*: $uv/\{1 - \theta(1 - u)(1 - v)\}$.
(2) *Clayton copula*: $(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$.
(3) *Frank copula*: $-\theta^{-1}\log([1 + (e^{-\theta u} - 1)(e^{-\theta v} - 1)/(e^{-\theta} - 1)])$.
(4) *Gumbel copula*: $\exp\{-((-\log u)^{\theta} + (-\log v)^{\theta})^{1/\theta}\}$.
(5) *Joe copula*: $1 - ((1 - u)^{\theta} + (1 - v)^{\theta} - (1 - u)^{\theta}(1 - v)^{\theta})^{1/\theta}$.
(6) *Plackett copula*: $[1 + (\theta - 1)(u + v) - \{\{1 + (\theta - 1)(u + v)\}^2 - 4\theta(\theta - 1)uv\}^{1/2}]/\{2(\theta - 1)\}$.

These copulas cover a very wide range of distributional shapes. The MLE and IFM estimator that are used in this simulation study are those that correspond to the case when the marginal distributions are assumed to be normal. To evaluate the robustness properties, the following sets of marginal distributions are considered:

(1) X_1 and X_2 are normally distributed (no misspecification in this case).
(2) $X_1 \sim t_\nu$ and $X_2 \sim t_\nu$.
(3) $X_1 \sim t_\nu$ and $X_2 \sim$ skew t-distribution (df $= \nu$, skewness $= 0.5$).
(4) $X_1 \sim t_\nu$ and $X_2 \sim \chi^2_2$.

The first corresponds to the correct specification of the marginal distributions, while each of the others leads to a misspecification of the model for ML and IFM. A skew t-distribution (df $= \nu$, skewness $= 0.5$) has tails that are of the same order as that for t_ν but the probability masses on either sides of the origin are different, leading to skewness (for example, see Bauwens and Laurent, 2005). The values 3 and 8 were considered for the degrees of freedom, ν.

Since the SP method estimates each marginal distribution nonparametrically, it is meant to be used when the sample size is moderate to large. In this study, we considered the sample sizes 40, 100, and 500. These three values capture a broad range of realistic settings, including financial studies where the sample sizes are usually quite large. Three values of θ were considered corresponding to the Kendall's τ of 0.2, 0.5 and 0.8, except for the AMH copula because its τ cannot take the value 0.5 or more. Therefore, the values of θ that we chose for the AMH copula are -0.5, 0.4, and 0.71, which correspond to $\tau = -0.1$, 0.1 and 0.2, respectively. These three values of θ for the AMH copula cover the parameter space well because $-1 < \theta < 1$.

All the computations were programmed in MATLAB Version 7.0.4. Optimizations were performed using the procedure fmincon.m in the Optimization Toolbox (3.0.2).

3.2. Results of the simulation study

Only a selection of the simulation results are presented here. Since the computations turned out to be time consuming, we restricted the number of repeated-samples to 250. The results are presented in Tables 1–6 and Figs. 1 and 2.

For a given method of estimation, let $\theta^{(i)}$ denote the estimator of θ for the ith repeated sample, θ_0 denote the true value, and N denote the number of repeated samples. To report the simulation results, let us introduce the following: *estimated bias* $= N^{-1}\sum \theta^{(i)} - \theta_0$, *estimated MSE* $= N^{-1}\sum\{\theta^{(i)} - \theta_0\}^2$, *estimated MSE-efficiency of a given estimator relative to IFM* $=$ {*estimated MSE of IFM*}{*estimated MSE of the estimator*}$^{-1}$. The last of these three quantities are tabulated in Tables 1, 4 and 5.

The main observations of the simulation study are summarized in the following subsections.

Table 1
MSE-efficiency (%) of the estimators relative to the IFM estimator when both marginal distributions are correctly specified as normal

	AMH			Clayton			Frank		
θ	−0.50	0.40	0.71	0.50	2.00	8.00	1.86	5.74	18.2
$\tau(\theta)$	−0.1	0.1	0.2	0.2	0.5	0.8	0.2	0.5	0.8
BM	102	105	108	116	167	237	103	107	116
ML	115	99	100	97	94	92	102	97	102
PML	95	101	98	72	78	90	101	91	88
	Gumbel			Joe			Plackett		
θ	1.25	2.00	5.00	1.44	2.86	8.77	2.50	11.6	115
$\tau(\theta)$	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
BM	117	150	199	117	172	232	109	112	108
ML	99	96	104	98	93	145	97	96	98
PML	81	83	94	79	89	83	88	101	95

The entries for BM, MLE, and PML are $\{MSE(IFM)/MSE(E)\} \times 100$, where E is BM, ML, and PML, respectively.
Number of repeated samples = 250; Number of observations in each sample = 100.

Table 2
Estimated biases of the estimators when the marginal distributions are incorrectly assumed to be normal for ML and IFM methods

θ	(T_3, T_3)			(T_3, χ^2_2)			$(T_3, SkewT_3)$		
	MLE	IFM	PML	MLE	IFM	PML	MLE	IFM	PML
AMH copula									
−0.50	0.02	−0.08	0.02	0.05	−0.02	−0.01	0.08	−0.03	0.01
0.40	0.09	0.08	−0.03	0.13	0.11	−0.02	0.16	0.15	−0.02
0.71	0.08	0.07	−0.00	0.12	0.10	−0.02	0.11	0.10	−0.03
Clayton copula									
0.50	−0.11	−0.14	0.05	0.06	−0.04	0.08	0.18	0.00	0.05
2.00	0.50	−0.04	0.05	−0.25	−0.71	0.08	0.76	−0.35	0.09
8.00	1.74	−0.88	−0.18	−3.97	−5.54	−0.21	−1.15	−4.40	−0.38
Frank copula									
1.86	0.78	0.65	0.01	0.65	0.45	0.06	1.11	0.79	0.03
5.74	1.66	1.06	0.02	1.22	0.38	0.10	1.74	0.83	−0.02
18.2	1.72	−0.51	−0.36	−1.99	−5.36	−0.31	0.78	−3.23	−0.15
Gumbel copula									
1.25	0.02	0.00	0.03	−0.08	−0.07	0.04	0.01	−0.05	0.04
2.00	0.24	0.06	0.07	−0.14	−0.25	0.04	0.01	−0.16	0.06
5.00	0.66	0.15	0.02	−1.21	−1.98	−0.03	−0.49	−1.40	0.03
Joe copula									
1.44	−0.01	−0.06	0.05	−0.20	−0.21	0.05	−0.16	−0.19	0.04
2.86	0.52	0.06	0.07	−0.28	−0.68	0.04	0.10	−0.50	0.07
8.77	0.58	−0.86	−0.34	−1.44	−4.75	−0.28	−0.45	−3.96	−0.36
Plackett copula									
2.50	1.79	1.22	0.22	1.19	0.66	0.19	1.46	0.90	0.05
11.6	8.33	4.05	0.73	3.58	−0.03	0.60	7.27	2.14	0.45
115	29.4	1.76	−5.18	−42.3	−74.6	−6.10	−10.8	−57.0	−11.0

Number of repeated samples = 250, and number of observations in each sample = 100.

3.2.1. Each marginal distribution is correctly specified as normal for ML and IFM methods

The results are given in Table 1. Since the marginal distributions and the copula are correctly specified, there is no mis-specification. As expected, the ML and the IFM estimators perform better than the PMLE. However, the differences are not large. Further: (i) the IFM and the ML estimators are equally good, and (ii) as expected, BM performed better than the other three estimators and its MSE relative to IFM and ML ranged up to 240%. Thus, as a result of the mean and variance of the marginal distribution being unknown, the ML and IFM methods suffer loss of efficiency. Consistency of the estimators is reflected in the mean of the estimators in the simulation being close to the true value, in every case—these are not shown in the tables but are available from the authors.

3.2.2. Each marginal distribution is incorrectly specified as normal for ML and IFM methods

The results in Fig. 1 and Tables 2–5, are for the case when the parametric methods ML and IFM incorrectly assume that each of the marginal distributions is normal; hence these estimators may not be even consistent. The SP method assumes that the marginal distributions are continuous, but apart from that it does not assume any functional form for these distributions. Thus, in contrast to the ML and IFM methods, the SP method is not based on incorrect assumption for the marginal distributions. Further, it is known that the SP estimator is consistent and asymptotically normal (see Genest et al., 1995; Tsukahara, 2005).

Tables 2 and 3 provide the estimated bias and the estimated standard deviations of the estimates based on the 250 repeated samples. Table 4 provides the results of the same study in a more concise form. The setting for Table 5 is the same except that the sample size is 500, instead of 100.

Figs. 1 and 2 provide histograms of the simulated estimates of θ for the Clayton, Joe and Plackett copulas. These figures illustrate an important finding of this study quite succinctly. They show that: (i) the ML and IFM estimators

Table 3
Estimated standard deviations of the estimators when the marginal distributions are assumed to be normal for ML and IFM methods

θ	(T_3, T_3)			(T_3, χ^2_2)			$(T_3, SkewT_3)$		
	MLE	IFM	PML	MLE	IFM	PML	MLE	IFM	PML
AMH copula									
−0.50	0.32	0.35	0.32	0.32	0.35	0.32	0.39	0.40	0.35
0.40	0.28	0.27	0.24	0.35	0.34	0.22	0.35	0.34	0.24
0.71	0.13	0.13	0.14	0.16	0.17	0.16	0.14	0.15	0.16
Clayton copula									
0.50	0.28	0.21	0.20	0.42	0.32	0.19	0.50	0.31	0.20
2.00	1.24	0.62	0.43	0.71	0.51	0.43	0.90	0.64	0.39
8.00	1.92	1.86	1.29	1.16	1.15	1.15	1.76	1.19	1.28
Frank copula									
1.86	0.98	0.88	0.65	1.00	0.83	0.64	1.46	1.09	0.69
5.74	1.39	1.01	0.81	1.17	0.85	0.82	1.62	1.13	0.89
18.2	2.05	2.27	1.74	1.74	1.66	1.78	2.11	2.29	1.91
Gumbel copula									
1.25	0.21	0.16	0.11	0.10	0.10	0.11	0.64	0.20	0.11
2.00	0.46	0.27	0.20	0.24	0.22	0.21	0.44	0.30	0.20
5.00	1.16	0.95	0.60	0.53	0.53	0.63	1.22	0.57	0.61
Joe copula									
1.44	0.31	0.22	0.17	0.13	0.14	0.18	0.47	0.27	0.19
2.86	0.92	0.58	0.41	0.65	0.35	0.39	1.08	0.59	0.40
8.77	0.83	1.87	1.34	1.54	1.08	1.29	1.15	1.42	1.20
Plackett copula									
2.50	3.36	1.59	0.89	1.85	1.06	0.82	2.38	1.37	0.77
11.6	9.56	5.15	3.51	5.38	3.14	3.17	8.11	4.24	3.38
115	31.4	36.1	27.8	24.6	14.0	27.1	32.8	21.6	25.3

Number of repeated samples $= 250$, and number of observations in each sample $= 100$.

are highly nonrobust against misspecification of the marginal distributions and hence unreliable as estimators of θ, and (ii) the distribution of the SP estimator is centered around the true value of θ and is far superior to the ML and the IFM estimators of θ. To save on space, the histograms are not given for the other cases, however the relevant information could be gleaned from Tables 2–5. Let us point out that, in some cases, the ML and the IFM estimators of θ were centered at a point close to the true value of θ even when the marginal distributions were misspecified (see Fig. 1).

We recognize that the very large values for relative MSE in Tables 4 and 5 are not precise, but we presented them anyway because they successfully convey the message that the ML and IFM estimators could be biased and the standard deviation of the estimators could become relatively unimportant compared to the bias. In summary, the SP method is considerably better than the ML and IFM methods.

3.2.3. Coverage rates of a PMLE-confidence interval for θ

Table 6 shows that an approximate 95% confidence interval based on a normal approximation for the large sample distribution of $\tilde{\theta}$, has coverage rates close to 95% for sample size $\geqslant 40$. The probability of coverage does not depend on the marginal distributions and hence they are not explicitly mentioned in Table 6. The coverage rate was at least 91% when $n \geqslant 100$, but it dropped to 88% when $n = 40$ for the Clayton copula with a large parameter value. Of course, confidence intervals for θ, based on the MLE and IFM, would be unreliable because the estimators are likely to be inconsistent. Thus, we conclude that the PML method also offers a reliable and easy to compute large sample confidence interval for θ.

Table 4
MSE-efficiencies (%) of MLE and the PMLE with respect to the IFM estimator

θ	MLE			PML		
	(T_3, T_3)	$(T_3, SkewT_3)$	(T_3, χ^2_2)	(T_3, T_3)	$(T_3, SkewT_3)$	(T_3, χ^2_2)
AMH copula						
−0.50	122	103	118	130	136	119
0.40	93	91	88	140	240	250
0.71	95	101	95	107	126	145
Clayton copula						
0.50	70	34	57	150	230	250
2.00	22	38	134	210	320	400
8.00	63	470	190	250	1170	2340
Frank copula						
1.86	76	54	62	290	380	210
5.74	46	35	30	330	250	130
18.2	76	310	450	170	430	970
Gumbel copula						
1.25	62	10	90	230	310	106
2.00	29	60	137	170	260	240
5.00	52	130	240	260	600	1070
Joe copula						
1.44	55	44	106	168	280	170
2.86	30	50	118	200	360	380
8.77	412	1160	540	220	1130	1360
Plackett copula						
2.50	28	34	32	480	450	220
11.6	27	19	24	330	195	95
115	71	310	240	164	490	750

Number of repeated samples = 250, and number of observations in each sample = 100.

4. An empirical example

In this section, we use a bivariate example to illustrate the estimation methods studied in this paper and to mention some of the challenges. We consider data from the 1988–1989 household expenditure survey conducted by the Australian Bureau of Statistics. For simplicity, we shall restrict to households consisting of exactly two adults and two children. The total sample size is 745. Let X_1=proportion of expenditure on housing and X_2=(1—proportion of expenditure on food). We wish to estimate the joint distribution of X_1 and X_2.

One of the challenges that we face is the specification of a suitable copula. Since there are a large number of copulas, specifying one that would suit a particular case in practice is not easy. Therefore, a reasonable strategy is to consider different copulas and evaluate their goodness of fits. To this end, we start with the so-called Archimedean family of copulas that have attracted considerable interest because of its flexibility and simplicity. The diagnostic checks that we performed suggest that the Frank copula fits well and better than the others considered. These are discussed below.

An Archimedean copula has the form $C(u, v) = \phi^{-1}\{\phi(u) + \phi(v)\}$, where ϕ is a strictly decreasing smooth convex function. This family is known to capture a range of functional forms. Genest and Rivest (1993) developed a graphical method to guide the choice of a suitable member of this family. In this method, we start with the definition, $\lambda(t) = \phi(t)/\{(\mathrm{d}/\mathrm{d}t)\phi(t)\}$. Since there is a one–one correspondence between the functions λ and ϕ, choosing the particular Archimedean copula is equivalent to choosing the function λ. A nonparametric estimator of λ is given by $\hat{\lambda}(t) = t - n^{-1}\sum I(t - V_i)$, where I is the indicator function and V_i = Number of $\{(X_{1j}, X_{2j}) : X_{1j} < X_{1i}, X_{2j} < X_{2i}\}/(n-1)$; see Genest and Rivest (1993). Their graphical procedure starts by sketching the λ functions for a range of Archimedean copulas and the estimate $\hat{\lambda}$. Now, a graphical way of choosing an Archimedean copula is to choose the one for which

Table 5
MSE-efficiencies (%) of MLE and the PMLE with respect to the IFM estimator

θ	MLE			PML		
	(T_3, T_3)	$(T_3, SkewT_3)$	(T_3, χ^2_2)	(T_3, T_3)	$(T_3, SkewT_3)$	(T_3, χ^2_2)
AMH copula						
−0.50	123	116	97	360	220	132
0.40	88	82	73	460	1070	840
0.71	89	87	89	230	590	940
Clayton copula						
0.50	14	40	65	680	890	760
2.00	18	32	240	410	1300	2750
8.00	59	170	190	700	7630	9500
Frank copula						
1.86	71	58	58	1840	1880	490
5.74	54	30	16	1980	1220	330
18.2	41	240	620	490	1170	4600
Gumbel copula						
1.25	8	7	44	280	470	610
2.00	7	11	48	520	640	860
5.00	12	133	250	590	2660	6260
Joe copula						
1.44	3	5	76	210	700	750
2.86	3	140	100	320	1050	1680
8.77	166	1820	580	570	5230	10350
Plackett copula						
2.50	52	26	44	2380	3430	740
11.6	27	14	6	1570	1140	115
115	44	540	290	680	1700	4500

Number of repeated samples = 250, and number of observations in each sample = 500.

Table 6
The coverage rates (in %) for a large sample 95% confidence interval for θ

Copula	θ	Sample size		
		40	100	500
AMH	{−0.50, 0.40, 0.71}	(96, 95, 93)	(95, 94, 90)	(91, 92, 94)
Clayton	{0.5, 2.0, 8.0}	(95, 92, 88)	(95, 94, 92)	(93, 92, 93)
Frank	{1.86, 5.74, 18.2}	(95, 96, 97)	(95, 94, 95)	(92, 96, 96)
Gumbel	{1.25, 2.00, 5.00}	(98, 96, 94)	(91, 94, 95)	(94, 96, 97)
Joe	{1.44, 2.86, 8.77}	(95, 95, 90)	(94, 96, 92)	(99, 93, 98)

The numbers in the parenthesis are the coverage rates corresponding to the three values of θ.
Number of repeated samples = 250.

the corresponding λ function is 'close' to $\hat{\lambda}$. Fig. 3 shows the λ functions for Clayton, Joe and Frank copulas. The λ functions for the other Archimedean copulas considered in simulation are not shown because they were not close to $\hat{\lambda}$. This figure suggests that the Frank copula fits the data better than the others.

We fitted the Clayton, Joe and Frank copulas by the ML, IFM and PML methods. The estimates and various relevant quantities are presented in Table 7. To assess the goodness of fit, the domain $[0, 1] \times [0, 1]$ of the copula was first partitioned into 25 squares of equal grid-size. Two of the cells were merged because of small counts. Even though the large sample distribution of the chi-square statistic corresponding to the PML method is unknown, the numerical values

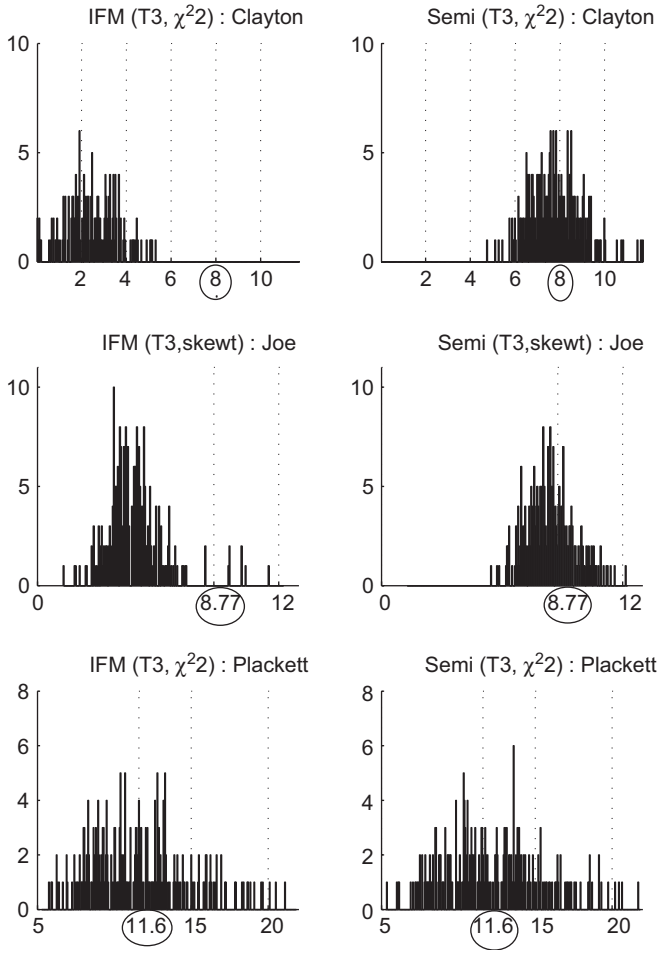

Fig. 1. Histogram for estimates of θ for Clayton, Joe and Plackett copula. The true values of θ are marked with circles. Number of repeated samples = 250, number of observations in each sample = 100. The two marginal distributions are given in the parentheses at the top of each histogram.

of the chi-square statistics can be used to compare the goodness of fit because they are all based on the same cells. If we were to make the assumption that the chi-square statistics corresponding to the PML methods were computed with the true marginal distributions, even though only the empirical distributions were used, then the large sample p-values, based on χ^2_{23}, would be 0.01, 0.09 and 0.40 for the Joe, Clayton and Frank copulas, respectively. These are given here just as a point of reference but we recognize that they do not have the usual meaning of the p-value. The p-values for the goodness of fit test statistic, S_n, of Genest et al. (2006) computed by the bootstrap method outlined in Section 4.2 therein, are also given in Table 7. These goodness of fit statistics suggest that only the Frank copula estimated by PML provides an acceptable fit.

A plot of X_1 vs X_2 is not that helpful in suggesting a suitable functional form for the copula. It is more helpful to plot $\{(\hat{F}_1(X_{1i}), \hat{F}_2(X_{2i})) : i = 1, \ldots, n\}$, where $\hat{F}_1$ and $\hat{F}_2$ are the empirical *cdf*s of the two variables; this is shown in Fig. 2. This figure shows that the points tend to concentrate near (0,0) and (1, 1). The shapes of the density functions of the estimated copulas are shown in Figs. 4–6. The single peaks for Clayton and Joe copulas do not appear to be consistent with the high concentration of points near the two vertices $\{(0, 0), (1, 1)\}$ in the unit square in Fig. 2. Thus, the Frank copula appears to be more appropriate.

Based on goodness of fit tests, sketches of the λ-functions, and comparison of the scatter plots with the shape of the copula density, the Frank copula estimated by PML appears to fit the data best. The PMLE of the Frank copula

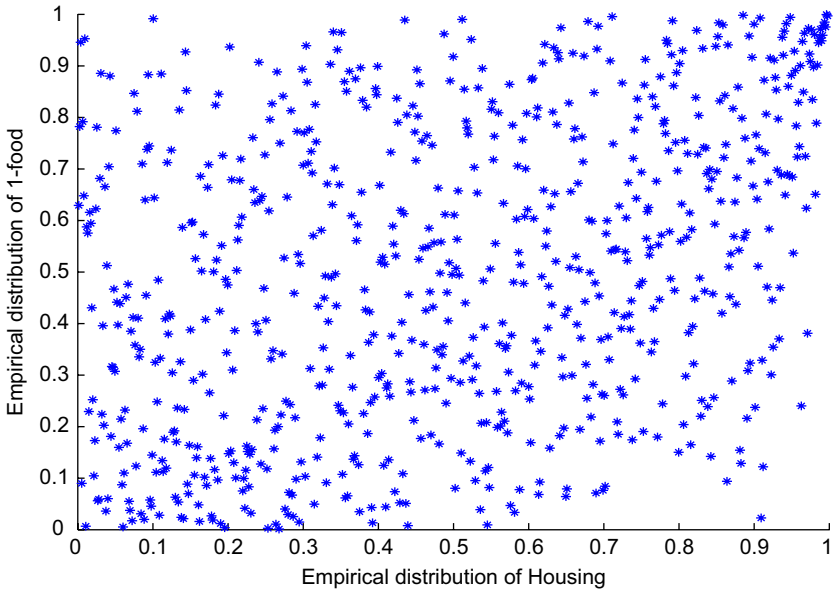

Fig. 2. Scatter diagram of $(\hat{F}_1(x_{1i}), \hat{F}_2(x_{2i}))$.

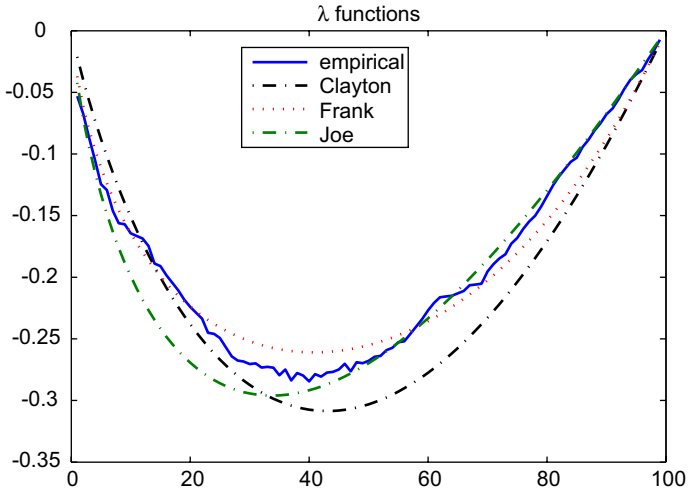

Fig. 3. The λ functions for different copulas and the empirical λ function.

parameter is 2.384 and its standard error, computed by Jack-knife, is 0.24. Therefore, the corresponding estimate of joint *cdf* of (X_1, X_2) is $C(\hat{F}_1(X_{1i}), \hat{F}_2(X_{2i}); 2.384)$, where C is the Frank copula defined earlier and $\hat{F}_j$ is the empirical *cdf* of F_j $(j = 1, 2)$.

The values of the copula parameter are difficult to interpret, but the corresponding values of the Kendall's τ and the Spearman ρ have more intuitive interpretations. The values of τ and ρ corresponding to the different estimates of θ are given in Table 7. Note that for the PML method, $\tilde{\theta} = 2.384$, $\tau(\tilde{\theta}) = 0.25$ and $\rho(\tilde{\theta}) = 0.37$. Apart from the intuitive interpretations of dependence conveyed by these estimates, the fact that $\tau(\tilde{\theta})$ and $\rho(\tilde{\theta})$ are almost identical to the nonparametric sample estimates, $\hat{\tau} = 0.25$ and $\hat{\rho} = 0.36$, respectively, suggest that the PML method captures these aspects of dependence well. This provides additional support to our previous observation that the Frank copula fits well and better than the other copulas.

Table 7
Parameter estimates and summary statistics for the household expenditure data

Copula	Estimate of θ (se)	Kendall's τ[b,c]	Spearman's ρ[b,c]	p-Value for χ^2-stat[d]	p-Value for GQR-test[e]
Empirical[a]		0.251	0.364		
PML					
Clayton	0.396 (0.04)	0.165	0.245	0.002	0.00
Frank	2.384 (0.24)	0.251	0.370	0.19	0.12
Joe	1.469 (0.07)	0.209	0.306	0.028	0.00
IFM					
Clayton	0.389 (0.06)	0.163	0.242	0.004	0.00
Frank	2.468 (0.25)	0.259	0.381	0.16	0.06
Joe	1.372 (0.05)	0.174	0.256	0.022	0.00
ML					
Clayton	0.419 (1.45)	0.173	0.257	0.003	0.00
Frank	2.505 (0.26)	0.263	0.386	0.15	0.04
Joe	1.417 (0.06)	0.190	0.279	0.025	0.00

[a]These are the nonparametric estimates of τ and ρ.
[b,c]The estimates of $\tau(\theta)$ and $\rho(\theta)$ corresponding to PML, IFM and ML were obtained by substituting the estimator of θ for θ.
[d]The chi-square goodness of fit statistics are all based on the same set of 24 cells and the p-values of the chi-square statistics are based on χ^2_{18}. The p-values corresponding to PML are given here as points of reference although their chi-square statistics may not be approximately χ^2_{18}.
[e]The GQR-statistic is the S_n statistic of Genest et al. (2006) with p-value computed by bootstrap.

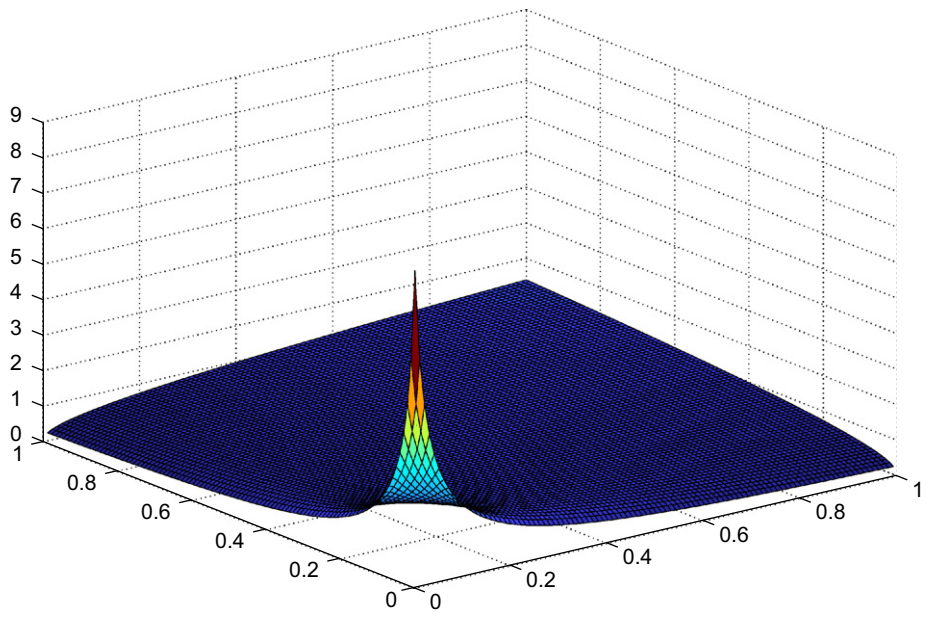

Fig. 4. Probability density function of the estimated Clayton copula.

The copula parameters estimated by the SP method and the parametric methods with the margins assumed to be normally distributed, turned out to be close (see Table 7). To illustrate the sensitivity of the IFM, we estimated the models with the standard normal replaced by the t_3-distribution for each margin. The IFM estimate of θ changed from $2.47(\text{se} = 0.25)$ to $1.83(\text{se} = 0.2)$, where the standard errors were computed by Jack-knife. Thus, the change in the parameter estimate as a result of changing the form of the marginal distribution is large, highlighting the nonrobustness of IFM.

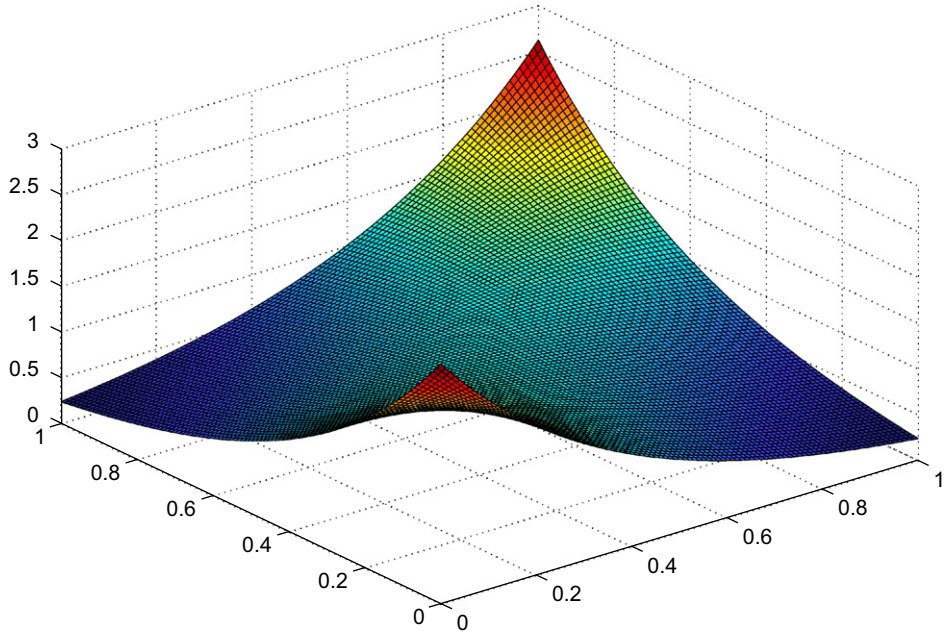

Fig. 5. Probability density function of the estimated Frank copula.

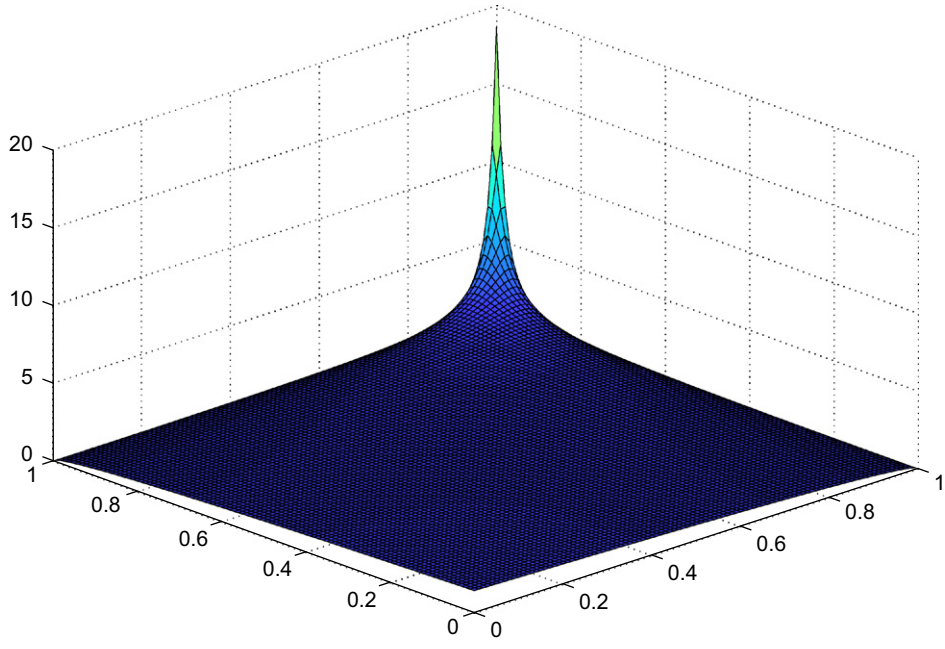

Fig. 6. Probability density function of the estimated Joe copula.

The method of estimating the joint distribution has a range of practical applications in modeling and data analysis. For example, if X_1 and X_2 are the returns from two investments, then the estimated *cdf* can be used to find the optimum investment portfolio of the form $aX_1 + (1 - a)X_2$, where $0 < a < 1$, that minimizes the probability $pr(aX_1 + (1 - a)X_2 < -K)$ of a large loss of magnitude K. Clearly, the method is also applicable to portfolios consisting of more than two investments. Such estimations are important in risk management.

5. Discussion

Copulas have been making significant inroads for modeling multivariate observations. Their main attraction is the flexibility that they offer to parameterize practically any shape for the distribution function. Copulas are particularly attractive in some areas because they specify the joint distribution in two stages enabling us to treat the univariate marginal distributions and the intrinsic joint behavior independently. The high degree of flexibility of copulas is associated with a whole range of potential approaches to using them in data analysis. Naturally, identifying the methods that are most suitable for a particular practical situation is of importance.

If the specification, $C\{F_1(x_1, \alpha_1), \ldots, F_k(x_k, \alpha_k); \theta\}$, of the copula model is correct then ML would be the preferred first option on the basis of its optimality in terms of asymptotic efficiency. However, over the years, IFM has emerged as preferable to ML. In view of the popularity of these two parametric methods, a closer examination of their suitability in practical situations is warranted. The main objective of this paper has been to provide evidence to show that, in most practical situations, it is better to use PML instead of IFM or ML.

Our simulations show that, when the marginal distributions are misspecified, the behavior of IFM is unpredictable and is quite likely to be inconsistent with very large MSE. By contrast, as expected, we find that PMLE is quite reliable. In terms of computing, PML is just as easy as IFM. The objective functions for estimating θ using IFM and PML are $C\{F_1(x_1, \hat{\alpha}_1), \ldots, F_k(x_k, \hat{\alpha}_k); \theta\}$ and $C\{\hat{F}_1(x_1), \ldots, \hat{F}_k(x_k); \theta\}$, respectively, and hence the statistical basis for these two objective functions are very close. Therefore, conceptually it is only a minor step to move between IFM and PML. The extent to which the PML performs better than the IFM, with hardly any additional costs in terms of computational requirements, strongly suggest that PML should be preferred to the popular IFM.

When using copulas for data analysis, the choice of appropriate copula is an important task. We have drawn from different sources to illustrate various diagnostics that may be useful in this regard.

It is possible to apply different SP methods for estimating θ, such as those based on minimum distance and estimating equations. A simulation study comparing such SP methods, may be found in Tsukahara (2005). It was observed there that the PMLE was one of the best two and that the difference between these two were small. Since our objective was to compare parametric with SP methods and PMLE has a well established literature, we restricted our study to PMLE.

In summary, the main observations of our study are:

(a) IFM and MLE are not robust against misspecification of the marginal distributions.
(b) The PML method is conceptually almost the same as the IFM but overcomes its nonrobustness against misspecification of the marginal distributions.
(c) In terms of statistical computations and data analysis, the PML is as easy to implement as the IFM.
(d) By using the PML method, one would not lose any important statistical insights as that would be gained by using IFM. An advantage of PML over IFM is that the former does not require modeling the marginal distributions explicitly.
(e) Our simulation results show that PMLE is better than ML and IFM estimators in most practical situations.

Thus, from a Computational Statistics and Data Analysis point of view, we argue that the PML method should be preferred to the IFM method.

6. Conclusion

Estimating the joint distribution of multivariate observations is of fundamental importance in statistical inference. The traditional parametric method involved assuming multivariate normal for the joint distribution and estimating the unknown parameters by MLE or other suitable methods. The limitation of this method has been well understood due to the limited range of distributional shapes that it can represent. The alternative method of using a completely nonparametric method, for example a multivariate kernel density estimate, would give a valid estimate for the density and distribution but one does not always wish to use a completely flexible method. The semiparametric (SP) method based on copulas discussed in this paper lies between these two extremes of completely parametric and completely nonparametric methods, and hence appears to offer an excellent compromise.

In this paper, we evaluated a SP method of estimating the copula. A simulation study showed that the SP method, which estimates the marginal distributions nonparametrically, is better than the fully parametric ML and IFM methods

when the marginal distributions are unknown. An example involving the household expenditure survey data, compared and contrasted the three methods. The example illustrated some of the difficulties in choosing the correct marginal distributions to implement fully parametric methods. By contrast, the SP method estimates the marginal distributions nonparametrically by the empirical distribution function (edf) and hence the difficult task of choosing the correct form for the marginal distribution does not arise.

In summary, the SP method of Genest et al. (1995) is close to the IFM method in approach, but the former is better than the latter when the marginal distributions are unknown, which is almost always the case. Further, this SP estimator and its standard errors are quite easy to compute. Thus, the evidence in favor of the SP method compared to the parametric methods, IFM and ML, is very strong indeed.

Acknowledgment

This research was partially supported by an Australian Research Council Discovery Project grant and a Monash University Postgraduate Student Award. The authors are grateful to the referees and Professor Christian Genest for many constructive suggestions.

References

Bauwens, L., Laurent, S., 2005. A new class of multivariate skew densities, with application to generalized autoregressive conditional heteroscedasticity models. J. Business Econom. Statist. 23, 346–354.

Cherubini, U., Luciano, E., Vecchiato, W., 2004. Copula Methods in Finance. Wiley, Chichester, UK.

Genest, C., MacKay, J., 1986. The joy of copulas: bivariate distributions with uniform marginals. Amer. Statist. 40, 280–283.

Genest, C., Rivest, L.-P., 1993. Statistical inference procedures for bivariate archemedean copulas. J. Amer. Statist. Assoc. 88, 1034–1043.

Genest, C., Werker, B.J.M., 2002. Conditions on the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models. In: Cuadras, C.M., Fortiana, J., Rodríguez-Lallela, J.A. (Eds.), Distributions with Given Marginals and Statistical Modelling. Kluwer, Dordrecht, The Netherlands, pp. 103–112.

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

Genest, C., Quessy, J.-F., Rémillard, B., 2006. Goodness-of-fit procedures for copula models based on the probability integral transformation. Scand. J. Statist. 33, 337–366.

Granger, C.W., Terasvirta, T., Patton, A.J., 2005. Common factors in conditional distributions for bivariate time series. J. Econometrics, in press.

Joe, H., 1997. Multivariate Models and Dependence Concepts. Chapman & Hall, London.

Joe, H., 2005. Asymptotic efficiency of the two-stage estimation method for copula-based models. J. Multivariate Anal. 94, 401–419.

McNeil, A.J., Frey, R., Embrechts, P., 2005. Quantitative Risk Management: Concepts, Techniques, and Tools. Wiley, NY.

Nelsen, R.B., 2006. An Introduction to Copulas. Springer, New York.

Patton, A.J., 2004. On the out-of-sample importance of skewness and asymmetric dependence for asset allocation. J. Finan. Econometrics 2 (1), 130–168.

Patton, A.J., 2005a. Estimation of multivariate models for time series of possibly different lengths. J. Appl. Econometrics 21, 147–173.

Patton, A.J., 2005b. Modelling asymmetric exchange rate dependence. Internat. Econom. Rev. 47, 527–556.

Shih, J., Louis, T., 1995. Inferences on the association parameter in copula models for bivariate survival data. Biometrics 51, 1384–1399.

Sklar, A., 1959. Fonctions de répartition à n dimensionset leurs marges. Publ. Inst., Statist. Univ. Paris 8, 229–231.

Tsukahara, H., 2005. Semiparametric estimation in copula models. Canad. J. Statist. 33, 357–375.

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