Efficient estimation of parameters in marginals in semiparametric multivariate models[*]

Valentyn Panchenko[†] Artem Prokhorov[‡]

March 18, 2010

Preliminary and incomplete – Please do not cite

Abstract

Recent literature on semiparametric copula models focused on the situation when the marginals are specified nonparametrically and the copula function is given a parametric form. For example, this setup is used in Chen, Fan and Tsyrennikov (2006) [Efficient Estimation of Semiparametric Multivariate Copula Models, JASA] who focus on efficient estimation of copula parameters. We consider a reverse situation when the marginals are specified parametrically and the copula function is modelled nonparametrically. This setting is no less relevant in applications. We use the method of sieve for efficient estimation of parameters in marginals and show its asymptotic distribution. Simulations suggest that the sieve MLE can be up to 40% more efficient relative to QMLE depending on the strength of dependence between marginals. An application using insurance company loss and expense data demonstrates empirical relevance of this setting.

JEL Classification: C13

Keywords: sieve MLE, copula, semiparametric efficiency

[*] Helpfull comments of seminar participants at University of Pittsburgh, University of New South Wales, Concordia University, QMF09 and FESAMES09 are gratefully acknowledged.

[†] Economics at the Australian School of Business, University of New South Wales, Sydney NSW 2052, Australia; email: valentyn.panchenko(at)unsw.edu.au

[‡] Department of Economics, Concordia University, Montreal, PQ H3G1M8 Canada; email: artem.prokhorov(at)concordia.ca

1 Introduction

Consider an m-variate random variable Y with joint pdf $h(y_1, \ldots, y_m)$. Let $f_1(y_1), \ldots, f_m(y_m)$ denote the corresponding marginal pdf's. Assume that the marginals are known up to a parameter vector β (β collects the distinct parameters of all marginals). The dependence structure is not given. We observe a random sample $\{\mathbf{y}_i\}_{i=1}^N = \{y_{1i}, \ldots, y_{mi}\}_{i=1}^N$. We are interested in estimating β efficiently without assuming anything about the joint distribution except for the marginals.

As a simple example consider the setting of a standard panel (small T, large N). We have a well specified marginal for each of T cross section (e.g., logit models, duration models, stochastic frontier models, etc.) and we are interested in efficient estimation of the parameters in the marginal distributions without assuming a parametric form on dependence between them. This setting is typical for microeconomic applications. The variable of interest y_t, $t = 1, \ldots, T$, can be the duration of unemployment in period t, or the use of social services in period t. Additional motivation for this problem comes from insurance. In particular, it arises in models of survival of multiple lives, where the two or more durations are dependent (see,e.g., Frees and Valdez, 1998). In life insurance of spouses this effect is known as the "broken heart" syndrome. In finance, a similar setting arises in the so called SCOMDY models (Chen and Fan, 2006a,b), where interest is in estimation of individual conditional distribution parameters and innovations of the univariate GARCH models are allowed to have arbitrary dependence.

We will use the well known representation of log-joint-density in terms of log-marginal-densities and the log-copula-density:

$$\ln h(y_1,\ldots,y_m;\beta)=\sum_{j=1}^{m}\ln f_j(y_j;\beta)+\ln c(F_1(y_1;\beta),\ldots,F_m(y_m;\beta)), \tag{1}$$

where $c(\cdots)$ is a copula density and F_i denotes the corresponding marginal CDF's. This decomposition is due to Sklar's (1959) theorem which states that any continuous joint distribution can be represented by a unique copula function of the corresponding continuous marginal CDF's.

It is well understood that the parameters of the marginals can be consistently estimated by maximizing the likelihood under the assumption of independence between marginals – this is the so called quasi maximum likelihood estimator, or QMLE. The copula term in (1) is zero in this case. However, QMLE is not efficient if marginals are not independent. For highly dependent marginals, the efficiency loss of QMLE relative to the full likelihood MLE may be quite large. In the context of a two-stage estimation of parametric copula models, Joe (2005) reports that FMLE asymptotic variance estimates for β are up to 93% smaller than those of QMLE. Recently, Prokhorov and Schmidt (2009) investigated the conditions for copula redundancy, that is when using the copula score does not improve efficiency over QMLE. The redundancy conditions they derive are fairly strong so incorporating information about dependence into parametric estimation problem will usually bring efficiency gains.

It is also well understood that, unlike QMLE, FMLE is generally not robust to copula misspecification. That is, the efficiency gains will come at the expense of an asymptotic bias if the joint density is misspecified. Prokhorov and Schmidt (2009) point out that there are robust parametric copulas, for which pseudo MLE (PMLE) based on an incorrectly specified copula leads to a consistent estimation. But a copula that is robust in one problem may not be robust in another, and some robust copulas are robust because they are redundant. So finding a general class of robust non-redundant parametric copulas is difficult if at all possible.

In this paper we address the issue of robust and efficient estimation of β using nonparametric methods. That is, we investigate whether we can obtain a consistent estimator of the parameters of marginals, which is more efficient than the QMLE, by modelling the copula nonparametrically. So the questions we ask are how to estimate β semiparametrically, what is the semiparametric efficiency bound for the estimation of β, and whether we can achieve it. To answer this questions we propose a sieve MLE (SMLE) procedure, which estimates β and $\ln c$ simultaneously (in one-step). Even though other nonparametric methods are available, e.g., kernel, local linear estimators, we choose the linear sieve method because of its simplicity. In effect we are replacing the true copula term in FMLE with its sieve approximator. Given the approximator, the problem becomes essentially identical to regular parametric FMLE. Subject to an approximation error, this produces a generally robust and usually non-redundant copula term, in the sense explained above.

What is not well understood is how such a semiparametric estimator compares to QMLE. Both QMLE and SMLE have the same parametric part – the correctly specified (up to a parameter vector) marginal distributions. They both involve no assumption on the dependence structure and, unlike FMLE, are robust to copula misspecification. However, it is unclear which estimator is more efficient. One the one hand, the SMLE uses dependence information and, as we shall show, is semiparametrically efficient. On the other, SMLE involves estimation of a number of extra parameters.

The paper is related to the literature on efficient semiparametric estimation of copula parameters with nonparametric marginals (see, e.g., Chen et al., 2006) and on efficient estimation of nonparametric marginals when the copula is fully known (see, e.g., Segers et al., 2008). More generally, it is related to the literature on seive-based estimation of models that contain unknown functions (see, e.g., Ai and Chen, 2003; Newey and Powell, 2003). It is also related to the literature on two-step semiparametric estimation (see, e.g., Newey and McFadden, 1994; Severini and Wong, 1992) and the literature on semiparametric efficiency bound (see, e.g., Severini and Tripathi, 2001; Newey, 1990).

The paper by Chen et al. (2006) considers a problem which is the converse of ours – a sieve MLE estimation when the copula has a known parametric form but the marginals are unknown. In that setting, sieves are employed to approximate univariate marginal densities. We are employing sieves to approximate a multivariate (log-)density. So the main difficulty of our setting is that, in high dimensions, we will suffer from the curse of dimensionality. For low dimensional problems, simulations show that SMLE is feasible and can lead to efficiency gains of up to 40% over QMLE.

We present theory of SMLE for our problem in Section 2. Section 3 contains simulation results, while Section 4 presents an insurance application. Section 5 contains concluding remarks.

2 Sieve MLE

Denote the true copula density by $c_o(\mathbf{u})$, $\mathbf{u} = (u_1, \dots, u_m)$, and denote the true parameter vector by β_o. Let $c_o(\mathbf{u})$ belong to an infinite-dimensional space $\Gamma = \{c(\mathbf{u}) : [0,1]^m \rightarrow [0,1], \int_{[0,1]^m} c(\mathbf{u}) = 1\}$ and β_o belong to $B \subset R^p$. Given a finite amount of data, optimization over the infinite-dimensional space Γ is not feasible. The method of sieves is used to overcome this problem. Define a sequence of approximating spaces Γ_N, called sieves, such that $\bigcup_N \Gamma_N$ is dense in Γ. Optimization is then restricted to the sieve space. Grenander (1981) is credited for observing that the MLE optimization, which is infeasible over an infinite dimensional space, is remedied if we optimize over a subset of the parameter space (i.e. over the sieve space) and then allow the subset to grow with the sample size. See Chen (2007) for a recent survey of sieve methods.

Chen (2007) suggests that a convenient finite dimensional linear sieve for approximating a multivariate log-pdf on $[0,1]^m$ is a tensor product of linear univariate sieves on $[0,1]$:

$$\Gamma_N = \left\{ c_{J_N}(\mathbf{u}) = \exp\left\{ \sum_{k=1}^{J_N} a_{1k} A_k(u_1) \cdot \ldots \cdot \sum_{k=1}^{J_N} a_{mk} A_k(u_m) \right\}, \right. \tag{2}$$

$$\left. \mathbf{u} \in [0,1]^m, \int_{[0,1]^m} c_{J_N}(\mathbf{u}) d\mathbf{u} = 1 \right\}, \tag{3}$$

$$J_N \rightarrow \infty \frac{J_N}{N} \rightarrow 0, \tag{4}$$

where $\{A_k\}$ contains known basis functions and $\{a_{jk}\}$ contains unknown sieve coefficients. Specific examples of the basis functions $A_k(u)$ include power series, trigonometric polynomials, splines, wavelets, neural networks and many others. For example, in simulations and application we use the trigonometric sieve basis functions:

$$A_k(u) = a_k \cos(k\pi u) + b_k \sin(k\pi u),$$

where $u \in [0,1]$ and $a_k, b_k \in R$. The number of sieve elements in the tensor sieve J_N^m is the smoothing parameter analogous to bandwidth in kernel estimation – it sets the quality of sieve approximation.

Since in general there is no analytic solution for the MLE of the sieve coefficients, the practical implementation of tensor sieves is often complicated. As an alternative we consider using Bernstein polynomials, in particular the Bernstein copula density introduced by Sancetta and Satchell (2004):

$$c_{J_N}(\mathbf{u}) = J_N^m \sum_{v_1=0}^{J_N-1} \cdots \sum_{v_m=0}^{J_N-1} \omega_v \prod_{l=1}^{m} \left(\begin{array}{c} J_N - 1 \\ \\ v_l \end{array} \right) u_l^{v_l} (1-u_l)^{J_N - v_l - 1}, \tag{5}$$

where ω_v denotes parameters of the polynomial indexed by $v = (v_1, \ldots, v_m)$ such that $0 \leq \omega_v \leq 1$ and $\sum_{v_1=0}^{J_N-1} \cdots \sum_{v_m=0}^{J_N-1} \omega_v = 1$. For the initial values of the parameters we may use the multivariate empirical density (histogram) estimator, i.e. $\omega_v = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(U_i \in H_v)$, where $U_i = (F_1(y_1), \ldots, F_m(y_m))$, $\mathbb{I}(\cdot)$ is the indicator function and

$$H_v = \left[\frac{v_1}{J_n}, \frac{v_1+1}{J_n}\right] \times \cdots \times \left[\frac{v_m}{J_n}, \frac{v_m+1}{J_n}\right]. \tag{6}$$

Note that the sieve above can be represented by a weighted sum of β-distributions. The relation between the empirical density and the MLE solution for ω still needs to be investigated but we found this sieve to converge faster in simulations than the tensor product sieve. Sancetta (2007) derives rates of convergence of the Bernstein copula to the true copula. Ghosal (2001), and references therein, discusses the rate of convergence of the sieve MLE based on Bernstein polynomial (only for one-dimensional densities.)

We can now write the sieve for $\Theta = B \times \Gamma$ as $\Theta_N = B \times \Gamma_N$. Further, let $\theta = (\beta', c)$, then the sieve MLE can be written as

$$\hat{\theta} = \arg\max_{\theta \in \Theta_N} \sum_{i=1}^{N} \ln h(\mathbf{y}_i; \theta) \tag{7}$$

This estimator is easy to implement – the estimation problem is in effect a parametric likelihood maximization problem once we replace Θ with Θ_N.

The parameter θ contains a parametric part that comes from the marginals β and a nonparametric part that describes the copula density c. We are interested in the asymptotic distribution of $\hat{\beta}$, the first p elements of $\hat{\theta}$. By the Gramér-Wold device, this distribution is normal if, for any $\lambda \in R^p, \|\lambda\| \neq 0$, the distribution of linear combination $\lambda'\beta$ is normal. Note that $\lambda'\beta$ is a functional of θ, call it $\rho(\theta)$. Its distribution given the sieve estimate $\hat{\theta}$ is known to depend on smoothness of the functional $\rho(\theta)$ and on the convergence rate of the nonparametric part of $\hat{\theta}$ (see, e.g., Shen, 1997). In our setting, the functional is very smooth and this will compensate for a slow convergence rate of the nonparametric part of $\hat{\theta}$ so that the parametric part of $\hat{\theta}$ will be $\sqrt{N}$-consistent.

In establishing asymptotic normality we follow the standard route (see, e.g., Ai and Chen, 2003; Chen et al., 2006). First, we show smoothness of $\lambda'\beta$ and then employ the Riesz representation theorem, to show normality of $\sqrt{N}\lambda'(\hat{\beta} - \beta)$. In showing semiparametric efficiency of the SMLE of β we follow the standard method of looking for the least favorable parametric submodel. A recent simplified version of this approach can be found in Severini and Tripathi (2001). In effect, by finding the Riesz representer we establish semiparametric efficiency exactly using that approach.

We first list standard identification and smoothness assumptions used in sieve based estimation (see, e.g., Chen, 2007; Chen et al., 2006).

Assumption 1 *(identification) $\beta_o \in$int$(B) \subset R^p$, B is compact and there exists a unique θ_o which maximizes $E[\ln h(\mathbf{Y}_i;\theta)]$ over $\Theta = B \times \Gamma$.*

A common smoothness assumption in nonparametrics is to restrict the class of considered functions by a certain smoothness property (see, e.g., Shen, 1997; Ai and Chen, 2003; Chen et al., 2006). Let g denote a real-valued, J times continuously differentiable function on $[0,1]^m$ whose J-th derivative satisfies the following condition for some $K > 0$ and $r \in (J, J+1]$:

$$|D^J g(x) - D^J g(y)| \leq K|x-y|_E^{r-J}, \text{for all } x, y \in [0,1]^m, \tag{8}$$

where $D^\alpha = \frac{\partial^\alpha}{\partial x_1^{\alpha_1} \ldots \partial x_m^{\alpha_m}}$, $\alpha = \alpha_1 + \ldots + \alpha_m$ is the differential operator, and $|x|_E = (x'x)^{1/2}$ is the Euclidean norm. Then g is said to belong to the Hölder class on $[0,1]^m$, denoted $\Lambda^r([0,1]^m)$. It is also called r-smooth on $[0,1]^m$. Linear sieves are known to approximate r-smooth functions well.

Assumption 2 *(smoothness - Hölder class for copula; differentiability for marginals)* $\Gamma =$

$\{c = \exp(g) : g \in \Lambda^r([0,1]^m), r > 1/2, \int c(u)du = 1\}$ *and* $\ln f_j(y_j;\beta), j = 1,\ldots,m$, *are twice differentiable w.r.t.* β

Now we introduce new notation that will be used in proofs of continuity of $\rho(\theta) = \lambda'\beta$ and of asymptotic normality and semiparametric efficiency of $\sqrt{N}(\hat{\beta} - \beta)$. First, we define the directional derivative of the loglikelihood in direction $\nu = (\nu_\beta', \nu_\gamma)' \in V$, where V is the linear span of $\Theta - \{\theta_o\}$,

$$\begin{aligned}
\dot{l}(\theta_o)[\nu] &\equiv \lim_{t\to 0} \frac{\ln h(\theta+t\nu,y)-\ln h(\theta,y)}{t}\Big|_{\theta=\theta_o} \\
&= \frac{\partial \ln h(\theta_o,y)}{\partial\theta'}[\nu] \\
&= \sum_{j=1}^m \left\{ \frac{\partial \ln f_j(y_j,\beta_o)}{\partial\beta'} + \frac{1}{c(F_1(y_1,\beta_o),\ldots,F_m(y_m,\beta_o))} \frac{\partial c(u_1,\ldots,u_m)}{\partial u_j}\Big|_{u_k=F_k(y_k,\beta_o)} \frac{\partial F_j(y_j,\beta_o)}{\partial\beta'} \right\} \nu_\beta \\
&\quad + \frac{1}{c(F_1(y_1,\beta_o),\ldots,F_m(y_m,\beta_o))} \nu_\gamma(u_1,\ldots,u_m)
\end{aligned}$$

Similarly, define

$$\begin{aligned}
\dot{\rho}(\theta_o)[\nu] &\equiv \lim_{t\to 0} \frac{\rho(\theta+t\nu)-\rho(\theta)}{t}\Big|_{\theta=\theta_o} \\
&= \lambda'\nu_\beta \\
&= \rho(\nu)
\end{aligned}$$

Then, we define the Fisher inner product $\langle\cdot,\cdot\rangle \equiv E\left[\dot{l}(\theta_o)[\cdot]\dot{l}(\theta_o)[\cdot]\right]$ on space V and the Fisher norm $||\nu|| \equiv \sqrt{\langle\nu,\nu\rangle}$, where expectation is with respect to the true density h. The closed linear span of $\Theta - \{\theta_o\}$ and the inner product $\langle\cdot,\cdot\rangle$ form a Hilbert space, call it $(\bar{V}, ||\cdot||)$.

Since $\rho(\theta) = \lambda'\beta$ is linear on $\bar{V}$ and $\dot{\rho}(\theta_o)[\nu] = \rho(\nu)$, to show smoothness of $\rho(\theta)$, we basically only need to establish that it is bounded on $\bar{V}$, i.e. that $\sup_{0\neq\theta-\theta_o\in\bar{V}} \frac{|\rho(\theta)-\rho(\theta_0)|}{||\theta-\theta_o||} < \infty$.

This will imply that $\rho(\theta)$ is continuous and its directional derivative is bounded as well, i.e. $\sup_{0\neq\nu\in\bar{V}} \frac{|\dot{\rho}(\theta_o)[\nu]|}{||\nu||} < \infty$. This is the case if and only if $\sup_{\nu\neq 0,\nu\in\bar{V}} \frac{|\lambda'\nu_\beta|^2}{||\nu||^2} < \infty$. So we now show when this condition holds.

Similar to Chen et al. (2006) and Ai and Chen (2003), we find the sup by writing

$$\begin{aligned} \sup_{\nu\neq 0,\nu\in\bar{V}} \frac{|\lambda'\nu_\beta|^2}{||\nu||^2} &= \sup_{\nu\neq 0,\nu\in\bar{V}} \left\{ |\lambda'\nu_\beta|^2 \left(E\left[\dot{l}(\theta_o)[\nu]^2 \right] \right)^{-1} \right\} \\ &= \lambda' \left(ES_\beta S_\beta' \right)^{-1} \lambda, \end{aligned} \tag{9}$$

where

$$\begin{aligned} S_\beta' &= \sum_{j=1}^m \left\{ \frac{\partial \ln f_j(y_j,\beta_o)}{\partial \beta'} + \left(\frac{1}{c(\mathbf{u})} \frac{\partial c(u_1,\ldots,u_m)}{\partial u_j} \right)\Bigg|_{u_k=F_k(y_k,\beta_o)} \frac{\partial F_j(y_j,\beta_o)}{\partial \beta'} \right\} \\ &\quad + \frac{1}{c(F_1(y_1,\beta_o),\ldots,F_m(y_m,\beta_o))} g^*(u_1,\ldots,u_m) \end{aligned} \tag{10}$$

and $(g_1^*,\ldots,g_p^*)$, which belong to the product space of square integrable zero-mean functions on $[0,1]^m$, are the solutions to the following infinite-dimensional optimization problem for $q = 1,\ldots,p$:

$$\begin{aligned} \inf_{g_q} E &\left[\sum_{j=1}^m \left\{ \frac{\partial \ln f_j(y_j,\beta_o)}{\partial \beta_q} + \left(\frac{1}{c(\mathbf{u})} \frac{\partial c(\mathbf{u})}{\partial u_j} \right)\Bigg|_{u_k=F_k(y_k,\beta_o)} \frac{\partial F_j(y_j,\beta_o)}{\partial \beta_q} \right\} \right. \\ &\quad \left. + \frac{1}{c(\mathbf{u})}\Bigg|_{u_k=F_k(y_k,\beta_o)} g_q(u_1,\ldots,u_m) \right]^2 . \end{aligned} \tag{11}$$

So the required condition is that $ES_\beta S_\beta'$ is finite and positive definite.

Assumption 3 *(nonsingular information) Assume that $ES_\beta S_\beta'$ is finite and nonsingular.*

Having established smoothness of $\rho(\theta)$ we can now appeal to the Riesz representation theorem (see, e.g., Kosorok, 2008, p. 328) to derive the asymptotic distribution of $\lambda'\beta$. Basically, the theorem states that for any continuous linear functional $L(\nu)$ on a Hilbert space there exists a vector ν^* (the Riesz representer of that functional) such that, for any ν

$$L(\nu) = \langle \nu, \nu^* \rangle,$$

and the norm of the functional defined as

$$||L||_* \equiv \sup_{||\nu|| \leq 1} ||L(\nu)||$$

is equal to $||\nu^*||$. The representer is used in derivation of normality and semiparametric efficiency of the sieve MLE.

Application of the theorem to $\dot{\rho}(\theta_o)[\nu] = \rho(\nu)$ suggests that there exists the Riesz representer $\nu^* \in \bar{V}$ such that $\lambda'(\hat{\beta} - \beta_o) = \langle \theta - \theta_o, \nu^* \rangle$ and $||\nu^*|| = \sup_{||\nu|| \leq 1} ||\rho(\nu)||$. The first claim implies that the asymptotic distributions of $\hat{\beta} - \beta_o$ and of $\langle \theta - \theta_o, \nu^* \rangle$ are identical – the latter is easier to use in proofs of normality than the former (see, e.g., Chen et al., 2006, Proof of Theorem 1). The second claim is used in proofs of semiparametric efficiency (Severini and Tripathi, 2001). Both are used to find the representer.

In fact we already found ν^* when we showed smoothness of $\rho(\theta)$ by finding $\sup_{\nu \neq 0, \nu \in \bar{V}} \frac{|\lambda' \nu_\beta|^2}{||\nu||^2}$. Since $\sup_{\nu \neq 0, \nu \in \bar{V}} \frac{|\lambda' \nu_\beta|^2}{||\nu||^2} = \sup_{||\nu|| \leq 1} ||\rho(\nu)||^2$, the representer is a vector whose norm, if squared, is equal to $\sup_{\nu \neq 0, \nu \in \bar{V}} \frac{|\lambda' \nu_\beta|^2}{||\nu||^2} = \lambda' \left(E S_\beta S_\beta' \right)^{-1} \lambda$. The vector is

$$\nu^* = \left(I, g^{*'} \right)' \left(E S_\beta S_\beta' \right)^{-1} \lambda$$

It is not difficult to show that

$$\begin{aligned} ||\nu^*||^2 &= E\left[\dot{l}(\theta_o)[\nu^*]\dot{l}(\theta_o)[\nu^*]\right] \\ &= \lambda' \left(ES_\beta S'_\beta\right)^{-1}\lambda, \end{aligned}$$

so the required condition holds.

The last assumption required for asymptotic normality is an assumption on the rate of convergence for the sieve MLE estimator of the unknown copula function. As in other sieve estimation literature, the sieve estimator is allowed to converge arbitrary slowly – smoothness of $\rho(\theta)$ compensates for that and the parametric part of the estimator is nevertheless $\sqrt{N}$ normal. For a discussion of convergence rates of different sieves see Chen (2007).

Assumption 4 *(convergence of sieve MLE) Assume that* $||\hat{\theta} - \theta_o|| = O_P(\delta_N)$ *for* $(\delta_N)^w = o(N^{-1/2})$, $w > 1$.

It is a variation of standard results (see, e.g. Shen, 1997; Severini and Tripathi, 2001; Chen et al., 2006) that, under these assumptions, $\hat{\beta}$ is consistent and the asymptotic variance of $N^{1/2}(\hat{\beta} - \beta)$ is equal to the semiparametric efficiency bound $||\nu^*||^2$.

Theorem 1 *Under Assumptions 1-4,* $\sqrt{N}(\hat{\beta} - \beta_o) \Rightarrow N(0, (E[S_\beta S'_\beta])^{-1})$ *and* $\hat{\beta}$ *is semiparametrically efficient.*

Given the consistent SML estimates $\hat{\beta}$ and $\hat{c}$, g_q^*'s can be estimated consistently in a sieve minimization problem as follows

$$\arg\min_{g_q \in \mathbf{A}_N} \left[\sum_{i=1}^{N} \sum_{j=1}^{m} \left\{ \frac{\partial \ln f_j(y_{ji}, \hat{\beta})}{\partial \beta_q} + \left(\frac{1}{\hat{c}(\hat{\mathbf{u}}_i)} \frac{\partial \hat{c}(\hat{u}_{1i}, \ldots, \hat{u}_{mi})}{\partial u_j} \right) \Bigg|_{\hat{u}_{ki} = F_k(y_{ki}, \hat{\beta})} \frac{\partial F_j(y_{ji}, \hat{\beta})}{\partial \beta_q} \right\} \right.$$
$$\left. + \sum_{i=1}^{N} \frac{1}{\hat{c}(F_1(y_{1i}, \hat{\beta}), \ldots, F_m(y_{mi}, \hat{\beta}))} g_q(\hat{u}_{1i}, \ldots, \hat{u}_{mi}) \right]^2,$$

where $q = 1, \ldots, p$ and $\mathbf{A}_N$ is one of the sieve spaces discussed above. Given consistent estimates $\hat{\beta}$, $\hat{c}$, and $\hat{g}^*$, a consistent estimate of $E[S_\beta S'_\beta]$ is easy to obtain if we replace the expectation evaluated at the the true values with a sample average evaluated at the estimates.

3 Simulations

Our initial simulations with linear tensor sieves, including splines, polynomials, and trigonometric polynomials, exhibit slow convergence rates. In contrast, using Bernstein polynomials, we were able to obtain the convergence within reasonable time. We therefore present the results for the latter sieve.

One of the practical problems we face is the choice of the degree of polynomials J_N in finite samples. While some asymptotic results on the rate of convergence and its dependence on J_N are available, they are not informative in the finite sample situation. The literature on sieves suggest using typical model selection techniques, such as BIC, AIC. However, the theoretical implications of using these techniques in the context of sieves are not explored. Similar to bandwidth selection in kernel estimation, it should be possible to devise data driven methods for choosing the number of elements in sieve but we do not pursue this point in this paper.

The DGP we use in simulations is similar to Joe (2005) who studied asymptotic relative efficiency (ARE) of likelihood based estimators, i.e. the ratio of asymptotic variance of Full MLE to that of QMLE of parameters in marginals. Joe (2005) finds that the ARE depends on the specification of marginals and copula. In particular, the higher is the dependence implied by the copula, the lower is the ARE of the QMLE, i.e. the more efficient is FMLE compared to QMLE. We take the case where the ARE is the lowest and investigate whether we may improve the efficiency of the QMLE by using the semiparametric sieve MLE technique.

We consider bivariate DGP with exponential marginals in which both mean parameters μ_1 and μ_2 are set to 0.5. The dependence is modelled by the Plackett copula with dependence parameter set equal to 0.002, which implies that we are close the lower Frechet bound for dependence. Joe (2005) reports ARE of 0.064 for QMLE of (μ_1, μ_2) in this specific case. In the simulation we use correctly specified marginals up to the two parameters to be estimated, while the copula function is modelled using the Bernstein polynomials sieve. We use the BIC to determine the degree of elements in the sieve J_N. The number of observations us $N = 1,000$.

Table 1 contains simulation results. BIC is minimized at $J_N = 8$. Thus we are estimating 64 nuisance parameters in the sieve and 2 parameters of the marginals. The optimization is complicated by the restrictions on sieve parameters and parameters of the marginals. We used standard constrained maximization routine in Matlab. Because of time constraints we used only 100 simulation runs. This will be extended in the future. We report the simulated mean of the Sieve MLE, QMLE and Full MLE estimators, their simulated variance and the simulated relative efficiency (RE) of the QMLE with respect to the Sieve MLE, i.e. the ratio of the SMLE simulated variance to that of QMLE.

Table 1: Simulated mean and variance for QMLE, SMLE, Plackett copula based FMLE

	μ_1 SMLE	QMLE	FMLE	μ_2 SMLE	QMLE	FMLE
$J_N = 8$						
Mean	0.488679	0.501630	0.499924	0.488633	0.498784	0.499509
Var	0.000126	0.000194	0.000012	0.000159	0.000233	0.000012
RE	0.649485			0.682403		
$J_N = 9$						
Mean	0.489800	0.501900	0.499951	0.489600	0.498300	0.500001
Var	0.000118	0.000194	0.000012	0.000153	0.000234	0.000012
RE	0.607530			0.653698		

The result suggests that in this specific situation we were able to improve the efficiency relatively to the QMLE substantially. The efficiency gain was as high as 32-40%. It appears that there is some evidence of downward bias in the estimates based on Sieve MLE for $J_N = 8$. Therefore, we try $J_N = 9$ in which case the bias seems to become smaller and the variance is also improved. This may suggest that BIC may not be the optimal procedure to select J_N. Note that this case corresponds to extremely high negative dependence between the marginals. In simulations using a weaker dependence, the improvements were not as substantial.

4 Application from insurance

We demonstrate the use of SMLE with an insurance application. We have data on 1,500 insurance claims. For each claim, we have the amount of claim payment, or loss, (Y_1) and the amount of claim-related expenses (Y_2). The claim-related expenses known as ALAE (allocated loss adjustment expense) include the insurance company expenses attributable to an individual claim, e.g. the lawyers' fees and claim investigation expenses. The claim amount variable is censored – there is a dummy variable, d, which is equal to one if a given claim has surpassed the policy limit and zero if not. For details of the data set, see Frees and Valdez (1998).

The claim amount and ALAE are assumed to be distributed according to the Pareto distribution with parameters (λ_1, θ_1) and (λ_2, θ_2), respectively:

$$F_j(Y_j) = 1 - \left(\frac{\lambda_j + Y_j}{\lambda_j}\right)^{-\theta_j}, \quad j = 1, 2. \tag{12}$$

Interest lies in efficient estimation of the marginal distribution parameters $(\lambda_1, \theta_1, \lambda_2, \theta_2)$, making efficient use of the strong dependence between the claim amount and ALAE. Additional complications arise due to censoring of Y_1. The likelihood contributions for censored observations will not be the same as for the uncensored ones and we need to account for that.

Define the marginal pdfs $f_j(y_j), j = 1, 2$. The QMLE log-likelihood contribution of an uncensored observation is $\ln f_j(y_j), j = 1, 2$. For a censored observation, the contribution is $\ln(1 - F_1(y_1)) = \theta_1(\ln(\lambda_1) - \ln(\lambda_1 + y_1))$. So for QMLE, the log-likelihood contribution of claim i is

$$l_i^Q = (1 - d_i) \ln f_1(y_{1i}) + d_i \ln(1 - F_1(y_{1i})) + \ln f_2(y_{2i}).$$

Now consider the joint likelihood. Define the joint cdf $H(y_1, y_2)$ and joint pdf $h(y_1, y_2)$. The FMLE contribution of an uncensored observation is $\ln h(y_1, y_2) = \ln f_1(y_1) + \ln f_2(y_2) + \ln c(F_1(y_1), F_2(y_2))$. To derive the contribution of a censored observation we follow Frees and Valdez (1998) in observing that $Prob(Y_1 \geq y_1, Y_2 \leq y_2) = F_2(y_2) - H(y_1, y_2)$. So the log-likelihood contribution of a censored observation is $f_2(y_2) - H_2(y_1, y_2)$, where $H_2(y_1, y_2) =$

$\frac{\partial H(y_1,y_2)}{\partial y_2}$. But $H(y_1,y_2) = C(F_1(y_1),F_2(y_2))$ so $H_2(y_1,y_2) = C_2(F_1(y_1),F_2(y_2))\, f_2(y_2)$, where $C_2(u_1,u_2) = \frac{\partial C(u_1,u_2)}{\partial u_2}$. Therefore the full log-likelihood contribution for observation i can be written as

$$\begin{aligned} l_i^F &= (1-d_i)[\ln f_1(y_1) + \ln f_2(y_2) + \ln c(F_1(y_1),F_2(y_2))] \\ &\quad + d_i[\ln f_2(y_2) + \ln(1 - C_2(F_1(y_1),F_2(y_2)))]. \end{aligned}$$

The main difficulty imposed by censoring is that we need to evaluate an additional term involving a copula derivative. For the SMLE, the term is approximated along with $\ln c$. For the FMLE, the term can be derived analytically for a given copula family or evaluated numerically.

The extra term will carry over to the variance problem (11) and a consistent estimate of the SMLE variance, $\hat{V}$, will now be

$$\begin{aligned} \arg\min_{g_q \in \mathbf{A}_N} \Bigg[& \sum_{i=1}^{N} (1-d_i) \Bigg\{ \sum_{j=1}^{2} \Bigg(\frac{\partial \ln f_j(y_{ji},\hat{\beta})}{\partial \beta_q} + \frac{1}{\hat{c}(\hat{\mathbf{u}}_i)} \frac{\partial \hat{c}(\hat{\mathbf{u}}_i)}{\partial u_j} \frac{\partial F_j(y_{ji},\hat{\beta})}{\partial \beta_q} \Bigg) + \frac{1}{\hat{c}(\hat{u}_{1i},\hat{u}_{2i})} g_q(\hat{u}_{1i},\hat{u}_{2i}) \Bigg\} \\ & + \sum_{i=1}^{N} d_i \Bigg\{ \frac{\partial \ln f_2(y_{2i},\hat{\beta})}{\partial \beta_q} - \frac{1}{1-\hat{C}_2(\hat{u}_{1i},\hat{u}_{2i})} \Bigg(\sum_{j=1}^{2} \frac{\partial \hat{C}_2(\hat{\mathbf{u}}_i)}{\partial u_j} \frac{\partial F_j(y_{ji},\hat{\beta})}{\partial \beta_q} + \int_0^1 g_q(s,\hat{u}_{2i})\, ds \Bigg) \Bigg\} \Bigg]^2, \end{aligned}$$

where $\beta = (\lambda_1,\theta_1,\lambda_2,\theta_2)'$, $\hat{u}_{ki} = F_k(y_{ki},\hat{\beta})$ and $q = 1,\ldots,4$. We will need to evaluate both g_q and its integral over u_1.

The three estimators, QMLE, FMLE and SMLE, and their standard errors are given in Table 2. The QMLE is an estimator based on the assumption of independence. It is known to be robust in the sense that it is consistent even if independence is a false assumption but to obtain the correct standard errors a “sandwich” formula for variance is needed. We report the robust standard errors in the table. The FMLE estimator is based on a fully specified parametric joint likelihood. We follow Frees and Valdez (1998) and assume the Frank copula with dependence parameter α, which along with the Pareto marginals completely parameterizes the model. Consistency of this estimator, sometimes called Pseudo-MLE, relies on correctness of the assumed copula family. If Frank is an incorrect copula family the FMLE results in a bias. The SMLE estimator is robust in the sense that it does not rely on a correctly specified parametric copula family. But it is not as efficient as any fully parametric model. So we should expect SMLE to be close to QMLE in terms of the estimates and to be between FMLE and QMLE in terms of standard errors.

Estimation results support this intuition. Our FMLE estimates using the Frank copula (which turn out almost identical to those in Frees and Valdez (1998)) provide evidence of an estimation bias that is not present in QMLE and SMLE, both of which are very close. This is an indication of robustness of QMLE and SMLE versus FMLE against a copula misspecification. While the FMLE standard errors are usually smaller than those of QMLE, which is an indication of higher efficiency – a compensation for the lack of robustness. The point we wish to stress is that the SMLE standard errors are smaller than those of QMLE and this gain comes at no robustness cost but at some computational cost. To obtain the SMLE, we used the cosine sieve with three elements in the sieve ($J_N = 3$). The choice was based on BIC.

Table 2: Estimates and standard errors of QMLE, SMLE, Frank copula based FMLE for insurance application

	QML Est. (Rob.St.Er.)	SML Est. (St.Er.)	FML Est. (St.Er.)
λ_1	14,442.57	14,438.91	14,561.68
	(2,385.31)	(1,434.87)	(1,392.08)
θ_1	1.135	1.136	1.115
	(0.127)	(0.067)	(0.065)
λ_2	15,133.78	15,133.78	16,708.93
	(2,172.04)	(1,549.66)	(1,833.18)
θ_2	2.223	2.223	2.312
	(0.246)	(0.142)	(0.188)
α			3.158
			(0.175)
LogL	-31,950.80	-31,813.60	-31,778.41

5 Concluding Remarks

We have proposed an efficient semiparametric estimator of marginal distribution parameters. This is a sieve maximum likelihood estimator based on a finite-dimensional approximation of the unspecified part of the joint distribution. As such, the estimator inherits the costs and benefits of the multivariate sieve MLE. The major benefit permitted by sieve MLE is the increased relative asymptotic efficiency compared to quasi-MLE. We showed that the efficiency gains are non-trivial. In some simulations the relative efficiency with respect to QMLE was about 0.6 – a 40% improvement.

The gains come at an increased computational expense. The MLE convergence is slow for the traditional sieves we considered. We found that the Bernstein polynomial is preferred to other sieves in simulations. The running times are greater than QMLE or full MLE assuming a parametric copula family but they are still reasonable (at least for the two dimensional problems we consider). Moreover, simulations reveal a downward bias in SMLE, which seems to be caused by the sieve approximation error – it decreases as the number of sieve elements increases.

A simple alternative to the proposed method is a fully parametric ML estimation problem. Although simpler computationally, it imposes an assumption on the dependence structure, which, if violated, renders the ML estimates inconsistent. In this respect, the semiparametric approach is more robust but clearly no more efficient than any parametric alternative.

References

Ai, C. and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, vol. 6, 5549–5632.

Chen, X. and Y. Fan (2006a): "Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification," *Journal of Econometrics*, 135, 125–154.

——— (2006b): "Estimation of copula-based semiparametric time series models," *Journal of Econometrics*, 130, 307–335.

Chen, X., Y. Fan, and V. Tsyrennikov (2006): "Efficient Estimation of Semiparametric Multivariate Copula Models," *Journal of the American Statistical Association*, 101, 1228–1240.

Frees, E. and E. Valdez (1998): "Understanding relationships using copulas," *North American Actuarial Journal*, 2, 1–25.

Ghosal, S. (2001): "Convergence rates for density estimation with Bernstein polynomials," *Annals of Statistics*, 29, 1264–1280.

Grenander, U. (1981): *Abstract Inference*, Wiley, New York.

Joe, H. (2005): "Asymptotic efficiency of the two-stage estimation method for copula-based models," *Journal of Multivariate Analysis*, 94, 401–419.

Kosorok, M. (2008): *Introduction to Empirical Processes and Semiparametric Inference*, Springer Series in Statistics, Springer.

Newey, W. and D. McFadden (1994): "Large sample estimation and hypothesis testing," .

Newey, W. K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58, 809–837.

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Prokhorov, A. and P. Schmidt (2009): "Likelihood-based estimation in a panel setting: robustness, redundancy and validity of copulas," *Journal of Econometrics*, 153, 93–104.

Sancetta, A. (2007): "Nonparametric estimation of distributions with given marginals via Bernstein-Kantorovich polynomials: L1 and pointwise convergence theory," *Journal of Multivariate Analysis*, 98, 1376–1390.

Sancetta, A. and S. Satchell (2004): "The Bernstein Copula And Its Applications To Modeling And Approximations Of Multivariate Distributions," *Econometric Theory*, 20, 535–562.

Segers, J., R. v. d. Akker, and B. Werker (2008): "Improving Upon the Marginal Empirical Distribution Functions when the Copula is Known," Discussion paper, Tilburg University, Center for Economic Research.

Severini, T. A. and G. Tripathi (2001): "A simplified approach to computing efficiency bounds in semiparametric models," *Journal of Econometrics*, 102, 23–66, doi: DOI: 10.1016/S0304-4076(00)00090-7.

SEVERINI, T. A. AND W. H. WONG (1992): "Profile Likelihood and Conditionally Parametric Models," *The Annals of Statistics*, 20, 1768–1802.

SHEN, X. (1997): "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591.

SKLAR, A. (1959): "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.