Efficient estimation of parameters in marginals in semiparametric multivariate models*

Valentyn Panchenko† Artem Prokhorov‡

Mar 2016

Abstract

We consider a general multivariate model where univariate marginal distributions are known up to a common parameter vector and we are interested in estimating that vector without assuming anything about the joint distribution, except for the marginals. If we assume independence between the marginals and maximize the resulting quasi-likelihood, we obtain a consistent but inefficient estimate. If we assume a parametric copula (other than independence) we obtian a full MLE, which is efficient but only under correct copula specification and badly biased if the copula is misspecified. Instead we propose a sieve MLE estimator which improves over QMLE but does not suffer the drawbacks of the full MLE. We model the unknown part of the joint distribution using the Bernstein-Kantorovich polynomial copula and assess the resulting improvement over QMLE and over misspecified FMLE in terms of relative efficiency and robustness. We derive the asymptotic distribution of the new estimator and show that it reaches the semiparametric efficiency bound. Simulations suggest that the sieve MLE can be almost as efficient as FMLE relative to QMLE provided there is enough dependence between the marginals. An application using insurance company loss and expense data demonstrates empirical relevance of the estimator.

JEL Classification: C13

Keywords: sieve MLE, copula, semiparametric efficiency

*Helpfull comments of seminar participants at University of Toronto, University of Pittsburgh, University of New South Wales, Concordia University, QMF, International Panel Data and FESAMES are gratefully acknowledged.

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

‡University of Sydney Business School, Sydney NSW 2006, Australia; email: artem.prokhorov(at)sydney.edu.au

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, where β collects all the distinct parameters in the marginals. The dependence structure between the marginals is not parameterized. We observe an i.i.d. sample $\{\mathbf{y}_i\}_{i=1}^N = \{y_{1i}, \ldots, y_{mi}\}_{i=1}^N$ and we are interested in estimating β efficiently without assuming anything about the joint distribution except for the marginals.

As an example consider the setting of a standard panel (small T, large N). We have a well specified marginal for each of T cross sections (e.g., logit models, duration models, stochastic frontier models, etc.) and we are interested in efficient estimation of the parameters in the marginal distributions with no apriori knowledge of the form or strength of dependence between them. This or similar setting is often encountered in microeconomic and actuarial applications (see, e.g, Winkelmann, 2012; Amsler et al., 2014; Frees and Valdez, 1998). In finance, a similar setting arises in the so called SCOMDY models and in other multivarariate GARCH-type models, where interest is in estimation of univariate conditional distribution parameters while the error terms are allowed to have arbitrary dependence (Chen and Fan, 2006a,b; Hafner and Reznikova, 2010).

However, recent literature on semiparametric copula models has focused on the case when the marginals are specified nonparametrically and the copula function is given a parametric form (see, e.g., Chen et al., 2006; Segers et al., 2008), which is an appropriate setting for many financial applications where it is important to parameterize dependence. In our setting, dependence is used solely to provide more precision in estimation of marginal parameters so we study the converse problem.

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 $F_i(\cdot)$ denotes the corresponding marginal cdf. 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 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 the marginals – this is the so called quasi maximum likelihood estimator, or QMLE. The copula term in (1) is zero in this case because the independence copula density is equal to one. However, QMLE is not efficient if marginals are not independent and for highly dependent marginals, the efficiency loss relative to the correctly specified full likelihood MLE is quite large. Joe (2005), for instance, reports up to 93% improvements in relative efficiency over QMLE in simulations when the full likelihood is correctly specified.

The situation when using copula terms in the likelihood does not improve asymptotic efficiency over QMLE is known as copula redundancy. Prokhorov and Schmidt (2009) derived a necessary and sufficient condition for copula redundancy and showed that such situations are very rare. Essentially, a parametric copula is redundant for estimation of parameters in the marginals if and only if the copula score with respect to these parameters can be written as a linear combination of the marginal scores – a condition generally violated for most commonly used parametric copula families and marginal distributions. As a result, significant efficiency gains remain unexploited.

An alternative that is more efficient asymptotically is a fully parametric estimation of the entire multivariate distribution by full MLE. This means assuming a parametric copula specification in addition to the marginal distributions. It is now 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 the pseudo MLE (PMLE) using an incorrectly specified copula family leads to a consistent estimation. However, copula robustness is problem specific and some robust copulas are robust because they are redundant. So finding a general class of robust non-redundant copulas remains an unresolved problem.

In this paper we address this problem using a semiparametric approach. That is, we

investigate whether we can obtain a consistent estimator of β, which is relatively more efficient than QMLE, by modelling the copula term nonparametrically. We use sieve MLE (SMLE) to do that. The questions we ask are whether a sieve-based copula approximator is the robust non-redundant alternative to QMLE and PMLE and what is the semiparametric efficiency bound for the SMLE of β. So our paper relates to the literature on sieve estimation (see, e.g., Ai and Chen, 2003; Newey and Powell, 2003; Bierens, 2014) and on semiparametric efficiency bounds (see, e.g., Severini and Tripathi, 2001; Newey, 1990).

The paper is organized as follows. In Section 2 we define our estimator and prove consistency, asymptotic normality and semiparametric efficiency. Section 3 contains simulation results, confirming the significant efficiency gains permitted by SMLE. Section 4 presents an insurance application. Section 5 contains concluding remarks.

2 Sieve MLE

Denote the true copula density by $c_o(\mathbf{u})$, $\mathbf{u} = (u_1, \ldots, u_m)$, and denote the true parameter vector by β_o. Let β_o belong to finite dimensional space $B \subset R^p$ and $c_o(\mathbf{u})$ belong to an infinite-dimensional space $\Gamma = \{c(\mathbf{u}) : [0,1]^m \to [0,1], \int_{[0,1]^m} c(\mathbf{u})d\mathbf{u} = 1, \int_{[0,1]^{m-1}} c_{J_N}(\mathbf{u}_{\text{-}\ell})d\mathbf{u}_{\text{-}\ell} = 1, \forall \ell\}$, where $\mathbf{u}_{\text{-}\ell}$ excludes u_ℓ. These conditions reflect that any copula is a joint probability distribution on the unit cube $[0,1]^m$ with uniform marginals. Given a finite amount of data, optimization over the infinite-dimensional space Γ is not feasible. The method of sieves is useful for overcoming this problem. Compared to other nonparametric methods such as kernels, local linear estimators, etc., the method of linear sieves is also quite simple – the infinite dimensional optimization is reduced to a regular parametric MLE.

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, known as the sieve space, and then allow the subset to grow with the sample size (see, e.g., Chen, 2007, for a survey of sieve methods).

There is a large number of convenient finite dimensional linear sieves known to work well for approximating univariate functions on $[0,1]$. To generalize them for multivariate copula setting, we can write them in a tensor product form as follows:

$$\Gamma_N = \Bigg\{ c_{J_N}(\mathbf{u}) = \sum_{k_1=1}^{J_N^{(1)}} \cdots \sum_{k_m=1}^{J_N^{(m)}} a_{k_1,\ldots,k_m} A_{k_1}(u_1) \times \ldots \times A_{k_m}(u_m) \, ,$$
$$\mathbf{u} \in [0,1]^m, \int_{[0,1]^m} c_{J_N}(\mathbf{u}) d\mathbf{u} = 1, \int_{[0,1]^{m-1}} c_{J_N}(\mathbf{u}_{\text{-}\ell}) d\mathbf{u}_{\text{-}\ell} = 1, \, \forall \ell \Bigg\},$$
$$J_N^{(\ell)} \to \infty, \, \frac{J_N^{(\ell)}}{N} \to 0, \, l = 1, \ldots, m$$

where $\{A_{k_\ell}\}$ are known univariate basis functions, $\{J_N^{(\ell)}\}$ is the number of basis elements in each direction ℓ and $\{a_{k_1,\ldots,k_m}\}$ are unknown sieve coefficients. Commonly used examples of basis functions $A_k(u)$ include power series, trigonometric polynomials, Fourier series, Chebyshev polynomials, splines, wavelets, neural networks and many others (see, e.g., Chen, 2007). The number of sieve elements in the tensor sieve $J_N^{(1)} \times \cdots \times J_N^{(m)}$ can be viewed as the smoothing parameter analogous to the bandwidth in a kernel estimation.

One of the challenges with the tensor product sieve is ensuring that the resulting sieve is the proper copula pdf, that is, it is non-negative, integrates to one and the marginals are uniform. Exponential or quadratic transformations are used often to ensure positivity and division by a normalizing constant is used to ensure that the sieve integrates to one (see, e.g., Chen et al., 2006). However, it is difficult to find an appropriate normalisation to ensure that all marginals are uniform. Moreover, the properties of the normalised objects, namely, the rates of convergence, may differ from the original sieve and may not be easy to derive. An alternative which does not require any transformation to satisfy the proper copula conditions is the Bernstein-Kantorovich polynomial (see, e.g., Sancetta and Satchell, 2004). In addition to this, the parameters of the Bernstein-Kantorovich sieve have a meaningful interpretation.

2.1 Bernstein-Kantorovich Sieve

The Bernstein-Kantorovich sieve is a tensor product sieve which uses β-densities as basis functions; it can be written as follows:

$$c_{J_N}(\mathbf{u}) = (J_N)^m \sum_{v_1=0}^{J_N-1} \cdots \sum_{v_m=0}^{J_N-1} \omega_{\mathbf{v}} \prod_{l=1}^{m} \begin{pmatrix} J_N - 1 \\ v_l \end{pmatrix} u_l^{v_l} (1-u_l)^{J_N - v_l - 1}, \tag{2}$$

where $\omega_{\mathbf{v}}$ denotes parameters of the polynomial indexed by multi-index $\mathbf{v} = (v_1, \ldots, v_m)$ such that $0 \leq \omega_{\mathbf{v}} \leq 1$ and $\sum_{v_1=0}^{J_N-1} \cdots \sum_{v_m=0}^{J_N-1} \omega_{\mathbf{v}} = 1$. These restrictions ensure that the above equation is a proper density. The interpretation of the coefficients $\omega_{\mathbf{v}}$ is that they are probability masses on an $J_N \times \cdots \times J_N$ grid (see, e.g., Zheng, 2011; Burda and Prokhorov, 2014).[1] In order to ensure that $c_{J_N^m}(\mathbf{u})$ is a copula density, i.e. that its marginals are uniform, we further require that $\sum_{\mathbf{v}_{\text{-}\ell}} \omega_{\mathbf{v}} = 1/(J_N)^{m-1}$, where multiple summations $\sum_{\mathbf{v}_{\text{-}\ell}}$ are performed over all elements of $\mathbf{v}$ except v_ℓ, $\ell = 1 \ldots m$.

The weights $\omega_{\mathbf{v}}$ can be viewed as a multivariate empirical copula density estimator, $\omega_{\mathbf{v}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(u_i \in H_{\mathbf{v}})$, where $u_i = (u_{i1}, \ldots, u_{im}) \in [0,1]^m$, $\mathbb{I}(\cdot)$ is the indicator function and

$$H_{\mathbf{v}} = \left[\frac{v_1}{J_N}, \frac{v_1+1}{J_N} \right] \times \cdots \times \left[\frac{v_m}{J_N}, \frac{v_m+1}{J_N} \right]. \tag{3}$$

So the Bernstein-Kantorovich polynomial sieve has the interpretation of a smoothed copula histogram where smoothing is done by the product of β-densities. Alternatively, it can be viewed as a mixture of the product of β-densities in u.

Sancetta (2007) derives the rates of convergence of the Bernstein-Kantorovich copula to the true copula. Petrone and Wasserman (2002) and Burda and Prokhorov (2014) established consistency of the Bernstein-Kantorovich polynomial when used as a prior on the space of densities on $[0,1]^m$ in a Bayesian framework. Ghosal (2001) and references therein discuss the rate of convergence of the sieve MLE based on the Bernstein polynomial (only for one-dimensional densities). Uniform approximation results for the univariate and bivariate Bernstein density estimator can be also found in Vitale (1975) and Tenbusch (1994). As $J_N \to \infty$, $c_{J_N}(\mathbf{u})$ is

[1] For simplicity we assume that J_N is the same for every dimension ℓ, but this assumption can be easily relaxed in cases where such asymmetry is required.

known to converge to the probability limit of the empirical copula estimator at every point on $[0,1]^m$ where the limit exists, and if it is continuous and bounded then the convergence is uniform (see, e.g., Lorentz, 1986).

This sieve is particularly attractive in our multivariate settings because of the uniform rate of convergence results available for c_{J_N} and because of the empirical copula interpretation of $\omega_{\mathbf{v}}$. The former ensures a relatively fast convergence compared to other tensor product sieves, which we observe in simulations, while the latter permits natural adaptive dimension reduction based on dropping $\omega_{\mathbf{v}}$'s which correspond to sparsely populated grid cells.

2.2 Asymptotic Properties

We can now write the sieve for $\Theta = B \times \Gamma$ as $\Theta_N = B \times \Gamma_N$, where Γ_N contains a generic vector of copula parameters γ. For the special case of the Bernstein-Kantorovich copula, $\gamma = \omega_{\mathbf{v}}$. Let $\theta = (\beta', c)$, then the sieve MLE (SMLE) can be written as follows

$$\hat{\theta} = \arg\max_{\theta \in \Theta_N} \sum_{i=1}^{N} \ln h(\mathbf{y}_i; \theta) \tag{4}$$

In essence, an infinite-dimensional problem over a space of functions is reduced to a finite-dimensional problem over a sieve of that space. As pointed out in sieve MLE literature (see, e.g., Chen, 2007), this estimator is very easy to implement in practice – it is a standard finite dimensional parametric MLE once we decide on the number of sieve copula coefficients, and, as we discuss later, a consistent estimator of the SMLE asymptotic covariance matrix can be obtained in some cases using standard MLE.

The initial θ vector is infinite dimensional because it contains the nonparametric part, $\ln c$, along with β. So the asymptotic distribution of $\hat{\beta}$ – the first p elements of $\hat{\theta}$ – depends on the behavior of $\hat{\theta}$ as its dimension grows. By the Gramér-Wold device, this distribution is normal if, for any $\lambda \in R^p, \|\lambda\| \neq 0$, the distribution of the linear combination $\lambda'\hat{\beta}$ is normal. Note that $\lambda'\beta$ is a functional of θ, call it $\rho(\theta)$. Given a sieve estimate $\hat{\theta}$, the asymptotic distribution of $\rho(\hat{\theta})$ depends on smoothness of the functional and on the convergence rate of the nonparametric part of $\hat{\theta}$ (see, e.g., Shen, 1997). In our setting, the functional is simple

and smooth. But the rate of convergence of the nonparametric part of $\hat{\theta}$ may be quite slow especially if m is large. It is a well established result in univariate settings that in such cases the smoothness of $\rho(\beta)$ compensates for this and a $\sqrt{N}$-convergence can be achieved for $\hat{\beta}$ (see, e.g., Bierens, 2014). We obtain a similar result in multivariate settings.

In establishing consistency and asymptotic normality we follow the standard route (see, e.g., Ai and Chen, 2003; Chen et al., 2006; Chen and Pouzo, 2009). 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 $\hat{\beta}$ we follow the standard method of looking for the least favorable parametric submodel. A simplified version of this approach can be found in Severini and Tripathi (2001). (In the proof of semiparametric efficiency in the Appendix we provide reference to that approach for readers more familiar with it.)

We now list identification and smoothness assumptions. Versions of these are commonly used in sieve estimation literature (see, e.g., Shen, 1997; Ai and Chen, 2003; Chen et al., 2006; Chen, 2007; Bierens, 2014).

Assumptions

A1 (identification) $\beta_o \in \text{int}(B) \subset R^p$, B is compact and there exists a unique θ_o which maximizes $E[\ln h(\mathbf{Y}_i;\theta)]$ over $\Theta = B \times \Gamma$.

A2 (smoothness) $\Gamma = \{c = \exp(g) : g \in \Lambda^r([0,1]^m), \int c(\mathbf{u})d\mathbf{u} = 1, \int_{[0,1]^{m-1}} c_{J_N}(\mathbf{u}_{\text{-}\ell})d\mathbf{u}_{\text{-}\ell} = 1, \forall \ell\}$, where $\Lambda^r([0,1]^m)$ denotes the Hölder class of r-smooth functions on $[0,1]^m$, $r > 1/2$, and $\ln f_j(y_j;\beta), j = 1,\ldots,m$, are twice continuously differentiable w.r.t. β.

The smoothness condition restricts log-copula-densities to the class of real-valued, continuously differentiable functions whose J-th order derivative satisfies Hölder's condition

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

where $D^\alpha = \frac{\partial^\alpha}{\partial x_1^{\alpha 1} \ldots \partial x_m^{\alpha m}}$ is the derivative operator, $\alpha = \alpha_1 + \ldots + \alpha_m$, $|x|_E = (x'x)^{1/2}$ is the Euclidean norm and K is a positive constant. All commonly used densities, including copulas, belong to this class, and various linear sieves, as well as the Bernstein-Kantorovich polynomial sieve, are known to approximate such functions well. In fact, commonly used copulas satisfy

the stronger property of Lipschitz continuity (see, e.g., Siburg and Stoimenov, 2008). But we use the more general smoothness property because it is common to nonparametric density estimation.

Let $\dot{l}(\theta_o)[\nu]$ denote the directional derivative, evaluated at θ_o, of the log-likelihood in direction $\nu = (\nu_\beta', \nu_\gamma)' \in V$, where V is the linear span of $\Theta - \{\theta_o\}$. Then,

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

where the last equation follows from (1). Similarly, define $\dot{\rho}(\theta_o)[\nu]$ as follows:

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

Let $\langle \cdot, \cdot \rangle$ denote the inner product based on the Fisher information metric on V and let $||\cdot||$ denote the Fisher information norm on V. Then, $\langle \nu_1, \nu_2 \rangle \equiv E\left[\dot{l}(\theta_o)[\nu_1]\dot{l}(\theta_o)[\nu_2]\right]$ and $||\nu|| \equiv \sqrt{\langle \nu, \nu \rangle}$, where expectation is with respect to the true density h. The closed linear span of $\Theta - \{\theta_o\}$ and the Fisher information metric form a Hilbert space, call it $(\bar{V}, ||\cdot||)$.

Since $\rho(\theta) = \lambda'\beta$ is linear on $\bar{V}$, in order to show smoothness of $\rho(\theta)$, we only need to establish that it is bounded on $\bar{V}$, i.e. that $\sup_{0\neq\theta-\theta_o\in\bar{V}} \frac{|\rho(\theta)-\rho(\theta_0)|}{||\theta-\theta_o||} < \infty$. Also, by the results in Shen (1997), boundedness of $\rho(\theta) = \lambda'\beta$ is necessary for $\rho(\theta) = \lambda'\beta$ to be estimable at the $\sqrt{N}$-rate. Boundedness of $\rho(\theta)$ will imply that $\rho(\theta)$ is continuous. Moreover, since $\dot{\rho}(\theta_o)[\nu] = \rho(\nu)$, boundedness of the directional derivative of $\rho(\theta)$ is equivalent to boundedness of $\rho(\theta)$ itself, i.e. it is equivalent to $\sup_{0\neq\nu\in\bar{V}} \frac{|\dot{\rho}(\theta_o)[\nu]|}{||\nu||} < \infty$. Because $\rho(\nu) = \lambda'\nu_\beta$, 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.

We follow Ai and Chen (2003) and Chen et al. (2006) and look for the minimal componentwise Fisher information metric for β. This minimization problem can be written as

follows:

$$\inf_{g_q} E\left[\sum_{j=1}^{m}\left\{\frac{\partial \ln f_j(y_j,\beta_o)}{\partial \beta_q}+\left(\frac{1}{c(\mathbf{u})}\frac{\partial c(\mathbf{u})}{\partial u_j}\right)\Big|_{u_k=F_k(y_k,\beta_o)}\frac{\partial F_j(y_j,\beta_o)}{\partial \beta_q}\right\}\right. \tag{5}$$
$$\left.+\left(\frac{1}{c(\mathbf{u})}g_q(u_1,\ldots,u_m)\right)\Big|_{u_k=F_k(y_k,\beta_o)}\right]^2,$$

where $E\left[\frac{1}{c(\mathbf{u})}g_q(\mathbf{u})\right]=0$. Let g_q^* denote the solution of (5), $q=1,\ldots,p$, and let $g^*=(g_1^*,\ldots,g_p^*)$.

We can now find the sup by writing

$$\begin{aligned}\sup_{\nu\neq 0,\nu\in\bar{V}}\frac{|\lambda'\nu_\beta|^2}{||\nu||^2} &= \sup_{\nu\neq 0,\nu\in\bar{V}}\left\{|\lambda'\nu_\beta|^2\left(E\left[\dot{l}(\theta_o)[\nu]^2\right]\right)^{-1}\right\} \\ &= \lambda'\left(ES_\beta S_\beta'\right)^{-1}\lambda,\end{aligned} \tag{6}$$

where

$$\begin{aligned}S_\beta' &= \sum_{j=1}^m\left\{\frac{\partial \ln f_j(y_j,\beta_o)}{\partial \beta'}+\left(\frac{1}{c(\mathbf{u})}\frac{\partial c(u_1,\ldots,u_m)}{\partial u_j}\right)\Big|_{u_k=F_k(y_k,\beta_o)}\frac{\partial F_j(y_j,\beta_o)}{\partial \beta'}\right\} \\ &\quad+\left(\frac{1}{c(\mathbf{u})}g^*(u_1,\ldots,u_m)\right)\Big|_{u_k=F_k(y_k,\beta_o)} \\ g^* &= (g_1^*,\ldots,g_p^*) \quad \text{and} \quad E\left[\frac{1}{c(\mathbf{u})}g_q^*(\mathbf{u})\right]=0.\end{aligned} \tag{7}$$

So $\rho(\theta)=\lambda'\beta$ in bounded if and only if $ES_\beta S_\beta'$ in (6) is a finite and positive definite matrix.

Assumption A3 (nonsingular information) Assume that $ES_\beta S_\beta'$ is finite and positive definite.

Having established smoothness of $\rho(\theta)$ we can use 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 will be used in the derivation of asymptotic normality and semiparametric efficiency of the sieve MLE.

The Riesz representation theorem, when applied to $\dot{\rho}(\theta_o)[\nu] = \rho(\nu)$, suggests that there exists a Riesz representer $\nu^* \in \bar{V}$ of $\rho(\nu)$, for which $\lambda'(\hat{\beta} - \beta_o) = \langle \hat{\theta} - \theta_o, \nu^* \rangle$ and $||\nu^*|| = \sup_{||\nu||\leq 1} ||\rho(\nu)||$. The first claim implies that the distributions of $\hat{\beta} - \beta_o$ and of $\langle \hat{\theta} - \theta_o, \nu^* \rangle$ are identical, which is useful for proving asymptotic normality of $\sqrt{N}(\hat{\beta} - \beta_o)$. The second claim is used in the proof of semiparametric efficiency. Both of these claims are useful for deriving the explicit form of the representer.

It turns out we have 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 $\sup_{\nu\neq 0, \nu\in\bar{V}} \frac{|\lambda'\nu_\beta|^2}{||\nu||^2} = \sup_{||\nu||=1} ||\rho(\nu)||^2$, the representer for our problem is a vector whose squared Fisher information norm is equal to $\sup_{\nu\neq 0, \nu\in\bar{V}} \frac{|\lambda'\nu_\beta|^2}{||\nu||^2} = \lambda' \left(E S_\beta S_\beta'\right)^{-1} \lambda$. It is straightforward to show that this vector can be written as follows

$$\nu^* = \left(I, g^{*'}\right)' \left(E S_\beta S_\beta'\right)^{-1} \lambda \tag{8}$$

As a check we can see that the squared Fisher information norm of ν^* can be written as follows

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

The last assumption required for asymptotic normality of $\sqrt{N}(\hat{\beta} - \beta_o)$ is an assumption on the rate of convergence for the sieve MLE estimator of the unknown copula function. As in other sieve literature, we allow the sieve estimator to converge arbitrary slowly – smoothness of $\rho(\theta)$ compensates for that and the parametric part of the estimator is still $\sqrt{N}$-estimable. We also impose a boundedness condition on the second order term in the Taylor expansion

of the sieve log-likelihood function. This technical condition will usually follow from the smoothness assumption **A2** but we state it explicitly to simplify the proof.

Assumption A4 (convergence of sieve MLE and smoothness of higher order term in Taylor expansion) Assume (A) that $||\hat{\theta}-\theta_o|| = O_P(\delta_N)$ for $(\delta_N)^w = o(N^{-1/2})$, $w > 1$ and there exists $\Pi_N\nu^* \in V_N - \{\theta_o\}$ such that $\delta_N||\Pi_N\nu^* - \nu^*|| = o(N^{-1/2})$ and (B) that, for any $\theta : ||\theta - \theta_o|| = O_p(\delta_N)$, the expected directional derivative $E\frac{d\dot{l}(\theta)[\nu]}{d\theta'}[\nu] \leq ||\nu||^2$.

A discussion of convergence rates of different sieves is provided by Chen (2007) and in references therein; general results on convergence rates of sieve MLE can be found in Wong and Severini (1991); Shen and Wong (1994). Basically, Assumption **A4** covers all commonly encountered sieves. For example, for the trigonometric sieve, Shen (1997) shows that $||\hat{\theta} - \theta_o|| = O_p(N^{-r/(2r+1)})$, where r is the Hölder exponent; Ghosal (2001) provides results on convergence rates of the Bernstein-Kantorovich sieve but using the Hellinger distance.

We can now state our main consistency and asymptotic efficiency results.

Theorem 1 *Under **A1**-**A4**, $\sqrt{N}(\hat{\beta} - \beta_o) \Rightarrow N(0, (E[S_\beta S_\beta'])^{-1})$.*

Proof. See Appendix for all proofs.

Theorem 2 *Under **A1**-**A4**, $||\nu^*||^2$ is the lower bound for semiparametric estimation of $\lambda'\beta$, i.e. $\hat{\beta}$ is semiparametrically efficient.*

In practice, one needs to estimate the asymptotic variance in order to conduct inference on β. The matrix $E[S_\beta S_\beta']$ can be estimated consistently as a sample average of $S_\beta S_\beta'$, once we obtain $\hat{\beta}$, $\hat{c}$, $\hat{g}_q^*$'s. Parameter estimates $\hat{\beta}$ and $\hat{c}$ are obtained in the sieve MLE but estimation of g_q^* requires a separate sieve minimization problem. We obtain consistent estimators g_q^* as solutions to the following problem

$$\arg\min_{g_q \in \mathbf{A}_N} \left[\sum_{i=1}^{N} \sum_{j=1}^{m} \left\{ \frac{\partial \ln f_j(y_{ji}, \hat{\beta})}{\partial \beta_q} + \left(\frac{1}{\hat{c}(\hat{\mathbf{u}}_i)} \frac{\partial \hat{c}(\hat{u}_{1i}, \ldots, \hat{u}_{mi})}{\partial u_j} \right) \Bigg|_{\hat{u}_{ki} = F_k(y_{ki}, \hat{\beta})} \frac{\partial F_j(y_{ji}, \hat{\beta})}{\partial \beta_q} \right\} \right. \quad (9)$$
$$\left. + \sum_{i=1}^{N} \frac{1}{\hat{c}(F_1(y_{1i}, \hat{\beta}), \ldots, F_m(y_{mi}, \hat{\beta}))} g_q(\hat{u}_{1i}, \ldots, \hat{u}_{mi}) \Big|_{\hat{u}_{ki} = F_k(y_{ki}, \hat{\beta})} \right]^2, \quad q = 1, \ldots, p$$

where $\mathbf{A}_N$ is one of the sieve spaces discussed above and $\hat{\beta}$ and $\hat{c}$ are consistent estimates of β and c and $\int g_q(\mathbf{u})/\hat{c}(\mathbf{u})d\mathbf{u} = 0$.

An alternative estimator of $E[S_\beta S_\beta']^{-1}$ was proposed by Ackerberg et al. (2012, 2014). It proceeds as follows. Using $\hat{\beta}$ and $\hat{c}$ we first evaluate the covariance matrix of all model parameters (both parameters in the marginal and in the copula) using the expected outer-product of the score. This is a large square matrix of dimension $p + J_N^m$. Then, we calculate the upper left $p \times p$ block of its inverse. Such estimator would be part of a standard MLE output. However, this method assumes that the likelihood is separable in β and c, which is not the case in our settings. This causes the estimate to be numerically unstable and so we use the sieve-based estimate above.

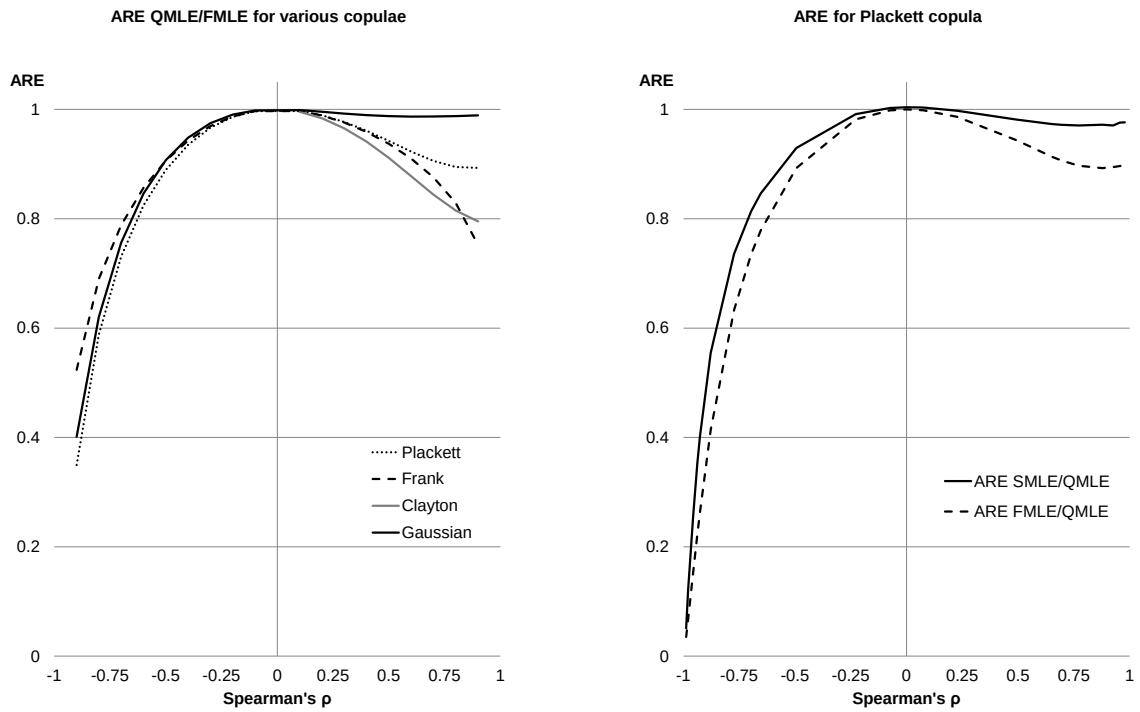
3 Simulations

Given the substantial advantage of Bernstein-Kantorovich polynomials in our setting we focus on sieves based on these polynomials.[2]. For Bernstein-Kantorovich sieves we observed robust convergence within reasonable time.

One of the practical problems we face is the choice of the degree of polynomial 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 suggests using typical model selection techniques, such as BIC and AIC or data driven methods such as cross-validation, so we use these methods.

The DGPs we use in the simulations are similar to those used by Joe (2005) who studies the asymptotic relative efficiency (ARE) of copula based MLE, i.e. the ratio of the asymptotic variance of FMLE to that of QMLE. Joe (2005) shows that ARE depends on the specification of marginals and copula as well as on the strength of dependence. Moreover, for some asymmetric marginal distributions, e.g., exponential, he finds that ARE for strongly nega-

[2]Our experimentations with linear tensor sieves, including splines, polynomials, and trigonometric polynomials, showed that these sieves are computationally intensive and exhibit slow convergence so they are not reported. However, a general matlab module implementing the Bernstein-Kantorovich and other sieves used in this and next sections is available at `http://research.economics.unsw.edu.au/vpanchenko/software/scopula.zip`



(a) (b)

Figure 1: Asymptotic relative efficiency of SMLE and FMLE.

tively dependent data is much larger than for positively dependent with the same dependence strength. Fig. 1 panel (a) reports the AREs as a function of dependence strength measured by Spearman's ρ for commonly used copulae, i.e., Gaussian, Clayton, Plackett, Frank. We report Spearman's ρ in the range from -0.9 to 0.9 as we experienced numerical instability in the region close to perfect negative or positive dependence for some copulae. Note that the Frank copula can accommodate only positive dependence. These plots confirm that there is a scope for improvement over the QMLE and that the largest gains can be expected in the case of strong negative dependence.

In our simulations we focus on a bivariate DGP with exponential marginals and the Plackett copula, which is comprehensive in the sense that it can accommodate the entire range of dependence captured by such measures as Kendall's τ or Spearman's ρ. The marginal dis-

tribution parameters (distribution means μ) are set at 0.5. The copula parameter takes 25 values in the range between 0.001 (Spearman's ρ near -1) and 1,000 (Spearman's ρ near +1).

Fig. 1 panel (b) reports the ARE for SMLE and FMLE as a function of Spearman's ρ when the data comes from the Plackett copula. The SMLE asymptotic variance is estimated using Eq. (9) for a sample of 200,000 observations, where we use the tensor product sieve with cosine basis functions without the intercept to approximate g_q. The number of sieve elements is $25 \times 25 = 625$.

Several observations are interesting in Fig. 1. First, as expected the AREs of both FMLE and SMLE are near one (subject to some estimation noise) in the case of independence, when we expect no gains over QMLE. Second, we observe the lowest ARE, that is the biggest efficiency gains, when Spearman's ρ approaches -1. This corresponds to extreme negative dependence and agrees with observations made by Joe (2005). In fact, the ARE of Plackett FMLE with copula parameter at 0.001 is 0.035 suggesting a 96.5% improvement over QMLE; the ARE of SMLE for the same strength of dependence is 0.051 suggesting a 94.9% improvement. Third, if we have strong positive dependence then FMLE does not show much efficiency gain over QMLE, which also agrees with Joe (2005). Interestingly, we observe a similar pattern for SMLE.

Joe (2005) provides an explanation for the asymmetry in ARE with respect to ρ. He observes that the upper and lower limits of dependence correspond to the Fréchet upper and lower bounds and he derives the constraints on the functional relationship between the two random variables implied by those bounds. Then, efficiency gains of FMLE over QMLE could be obtained only if the parameters of the marginals can be identified from these constraints. It turns out that for the bivariate distribution with exponential marginals under the Fréchet lower bound the parameters are identified and the ARE goes to zero in the limiting case of strong negative dependence. On the contrary, under the Fréchet upper bound the parameters of the marginals are not identified and the ARE is approaching one in the limiting case of positive dependence. The simulations summarized in Fig. 1 panel (b) show that similar arguments hold for SMLE.

Additionally, we ran the same analysis for different values of μ and non-exponential

marginals (not reported here). Interestingly, changing the true values of μ did not affect the ARE comparisons. This suggests that relative efficiency in estimation of parameters in exponential marginals depends only on the strength of dependence and not on the parameters in the marginals themselves. This isolates the effect of dependence on the ARE and makes exponential marginals particularly attractive for simulation purposes. For non-exponential marginals, the ARE denerally depends on both parameters in the marginals and the dependence parameter.

Naturally, the efficiency gains reported in Fig. (1b) using FMLE are higher than those obtained using SMLE. However, in regions close to extreme negative dependence and to independence the gap between the two estimators is diminishing. Also, it is perhaps surprising how close the semiparametric estimator is to a fully parametric one in terms of asymptotic precision. We stress that these results do not depend on the values of parameters in the marginals (other than that they must be identified).

Next we consider the performance of SMLE, FMLE and QMLE for a fixed value of Spearman's ρ. We keep exponential marginals and the Plackett copula as the DGP and set the true parameter value in the marginals at $\mu_1 = \mu_2 = 0.5$ and in the copula at $\gamma = 0.05$, which implies moderate negative dependence with Spearman's ρ of -0.77. The sample size is $N = 1,000$ and the number of simulations is $R = 1,000$.

Table 1 contains the simulation results. We report the mean value of the estimates for each marginal as well as various versions of the variance estimator and the MSE. Under *Var*, we report sample variance estimates while under *AVar*, we report estimates of the asymptotic variance obtained using a solution to (9). The number of elements in the Bernstein-Kantorovich sieve is $10 \times 10 = 100$. This number is picked by cross-validation. A key feature of the table is that SMLE shows substantial improvement over QMLE. The sample variance is close to the asymptotic variance.

Next we look at how the SMLE bias and variance change with the number of sieve elements. Table 2 reports means, variances and MSEs for the two estimates as well as the value of log-likelihood and three popular model selection criteria: leave-one-out likelihood cross-validation, AIC and BIC. The value of log-likelihood $LogL$ decreases as sieve complex-

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

	Parameter μ_1			Parameter μ_2		
	FMLE	SMLE	QMLE	FMLE	SMLE	QMLE
Mean	0.5004	0.4987	0.5001	0.4996	0.4976	0.4995
$N\times$ Var	0.1720	0.1919	0.2640	0.1740	0.1882	0.2590
$N\times$ MSE	0.1721	0.1937	0.2640	0.1741	0.1939	0.2592
$N\times$ AVar	0.1580	0.1860	0.2500	0.1580	0.1860	0.2500

Table 2: Optimal number of sieve elements in SMLE

J_n	Mean 1	Mean 2	Var 1	Var 2	MSE 1	MSE 2	CV crit	LogL	AIC	BIC
6	0.4968	0.4958	0.2012	0.2055	0.2116	0.2234	-640.46	188.33	448.65	625.33
7	0.4976	0.4965	0.1945	0.2005	0.2001	0.2126	-640.39	167.73	433.46	673.94
8	0.4982	0.4970	0.1918	0.1944	0.1952	0.2033	-640.00	152.15	432.30	746.40
9	0.4985	0.4974	0.1900	0.1908	0.1923	0.1976	-640.00	140.61	443.21	840.74
10	0.4987	0.4976	0.1919	0.1882	0.1937	0.1939	-639.96	131.76	463.52	954.30
11	0.4989	0.4977	0.1908	0.1904	0.1921	0.1958	-639.99	124.73	491.46	1085.30
12	0.4989	0.4977	0.1902	0.1907	0.1915	0.1961	-640.19	119.05	526.09	1232.81
13	0.4985	0.4977	0.1944	0.1938	0.1968	0.1992	-640.07	114.34	566.69	1396.10
14	0.4985	0.4975	0.1985	0.1962	0.2008	0.2022	-640.46	110.38	612.77	1574.69

ity grows as expected, while the information criteria AIC and BIC (especially BIC) seem to select an underparameterized model.

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:

$$F_j(Y_j) = 1 - \left(\frac{\lambda_j + Y_j}{\lambda_j}\right)^{-\theta_j}, \quad j = 1, 2. \tag{10}$$

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

Table 3: QMLE, SMLE, and FMLE for insurance claims and related expenses

	QML Est. (Rob.St.Er.)	Frank FML Est. (St.Er.)	G-H FML Est. (St.Er.)	SML Est.
λ_1	14,443.15	14,562.02	14,040.61	14,631.90
	(1,515.09)	(1,498.47)	(1,351.15)	(1,492.36)
θ_1	1.135	1.115	1.122	1.137
	(0.076)	(0.073)	(0.067)	(0.073)
λ_2	15,133.28	16,708.37	14,223.42	15,421.69
	(1,744.96)	(1,900.49)	(1,405.75)	(1,663.83)
θ_2	2.223	2.312	2.119	2.233
	(0.183)	(0.190)	(0.143)	(0.172)
α	-	3.158	-0.791	-
		(0.171)	(0.035)	
LogL	-31,951	-31,778	-31,749	-31,731

numerically.

The extra term will carry over to the variance problem (5) and a consistent estimate of the SMLE variance, $\hat{V}$, will now be

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

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 3. The QMLE 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.

The FMLE estimator is based on a fully specified parametric joint likelihood. We follow Frees and Valdez (1998) and assume the Frank and Gumbel-Hougaard copulas with dependence parameter denoted by α, which along with the Pareto marginals completely parameterize the model. Consistency of this estimator, sometimes called Pseudo-MLE, relies on

correctness of the assumed copula family. If Frank or Gumbel-Hougaard are incorrect copula families then the FMLE will be biased.

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. To obtain the SMLE, we use the Bernstein-Kantorovich sieve with $J_N = 10$ and to obtain the SMLE standard errors we use the cosine sieve with 9 parameters. The choice of the number of parameters in the Bernstein-Kantorovich sieve is based on cross-validation.

Estimation results support the above intuition. Our FMLE estimates using the Frank and Gumbel-Hougaard copula (which turn out virtually 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 supports the robustness (to copula misspecification) argument. The FMLE standard errors are usually smaller than those of QMLE. This indicates higher relative 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).

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. A major benefit is the increased precision compared to quasi-MLE, permitted by the use of dependence information. Simulations show that potential efficiency gains are huge. The efficiency bound is determined by the dependence strength and we show that our estimator reaches that bound.

The gains come at an increased computational expense. The convergence is slow for the traditional sieves we considered. We found that the Bernstein-Kantorovich polynomial is

preferred to other sieves. The running times are greater than the full MLE assuming an "off-the-shelf" parametric copula family but far from being prohibitive (at least for the two dimensional problem we consider). Moreover, simulations reveal a small 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. Moreover, robust parametric copulas are often robust because they are redundant. So the proposed estimator seems to offer a unique way of constructing a copula that is robust and generally non-redundant.

Methods to improve computational efficiency of SMLE focus on reducing the effective number of sieve parameters. Such methods involve penalized and restricted estimation and are particularly appealing for the Bernstein-Kantorovich polynomial where the sparse portions of the sieve parameter space correspond to histogram cells with little or no mass. We leave development of such methods for future work.

References

Ackerberg, D., X. Chen, and J. Hahn (2012): "A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators," *The Review of Economics and Statistics*, 94, 481–498.

Ackerberg, D., X. Chen, J. Hahn, and Z. Liao (2014): "Asymptotic Efficiency of Semiparametric Two-step GMM," *The Review of Economic Studies*.

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

Amsler, C., A. Prokhorov, and P. Schmidt (2014): "Using copulas to model time dependence in stochastic frontier models," *Econometric Reviews*, 33, 497–522.

Bierens, H. J. (2014): "Consistency and Asymptotic Normality of Sieve ML Estimators under Low-Level Conditions," *Econometric Theory*, FirstView, 1–56.

Burda, M. and A. Prokhorov (2014): "Copula based factorization in Bayesian multivariate infinite mixture models," *Journal of Multivariate Analysis*, 127, 200 – 213.

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.

Chen, X. and D. Pouzo (2009): "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals," *Journal of Econometrics*, 152, 46–60.

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.

Hafner, C. M. and O. Reznikova (2010): "Efficient estimation of a semiparametric dynamic copula model," *Computational Statistics & Data Analysis*, 54, 2609–2627.

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.

Lorentz, G. (1986): *Bernstein Polynomials*, University of Toronto Press.

Newey, W. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5, 99–135.

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

Petrone, S. and L. Wasserman (2002): "Consistency of Bernstein Polynomial Posteriors," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64, 79–100.

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," *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.

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

Shen, X. and W. H. Wong (1994): "Convergence Rate of Sieve Estimates," *The Annals of Statistics*, 22, 580–615.

Siburg, K. F. and P. A. Stoimenov (2008): "A scalar product for copulas," *Journal of Mathematical Analysis and Applications*, 344, 429 – 439.

Sklar, A. (1959): "Fonctions de repartition a n dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Universite de Paris*, 8, 229–231.

Stein, C. (1956): "Efficient Nonparametric Testing and Estimation," *Proceedings of Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 187–195.

Tenbusch, A. (1994): "Two-dimensional Bernstein polynomial density estimators," *Metrika*, 41, 233–253, metrika.

Vitale, R. (1975): "A Bernstein polynomial approach to density function estimation," in *Statistical inference and related topics*, ed. by M. Puri.

Winkelmann, R. (2012): "Copula bivariate probit models: with an application to medical expenditures," *Health economics*, 21, 1444–1455.

Wong, W. H. and T. A. Severini (1991): "On Maximum Likelihood Estimation in Infinite Dimensional Parameter Spaces," *The Annals of Statistics*, 19, 603–632.

ZHENG, Y. (2011): "Shape restriction of the multi-dimensional Bernstein prior for density functions," *Statistics and Probability Letters*, 81, 647–651.

6 Appendix

Proof of Theorem 1: Let $l_i(\theta) = \ln h(\mathbf{y}_i;\theta)$, $l(\theta) = \frac{1}{N}\sum_{i=1}^N l_i(\theta)$ and $0 < \varepsilon_N = o(N^{-1/2})$. We follow Chen et al. (2006) and consider a continuous path $\theta(t) = \hat{\theta} \pm t\varepsilon_N \Pi_N \nu^*$, $t \in [0,1]$, such that $\theta(0) = \hat{\theta}$ and $\theta(1) = \hat{\theta} \pm \varepsilon_N \Pi_N \nu^*$.

Under Assumption A2, $l(\theta)$ is twice continuously differentiable w.r.t. t and

$$\begin{array}{rcl} \left.\frac{d\,l(\theta(t))}{dt}\right|_{t=\tau} & = & \frac{1}{N}\sum_{i=1}^N \left.\frac{d\,l_i(\theta(t))}{dt}\right|_{t=\tau} = \frac{1}{N}\sum_{i=1}^N \frac{d\,l_i(\theta(\tau))}{d\theta'}\,[\pm\varepsilon_N\Pi_N\nu^*] \\ \left.\frac{d^2\,l(\theta(t))}{dt^2}\right|_{t=\tau} & = & \frac{1}{N}\sum_{i=1}^N \frac{d^2\,l_i(\theta(\tau))}{d\theta'd\theta}\,[\pm\varepsilon_N\Pi_N\nu^*, \pm\varepsilon_N\Pi_N\nu^*] \end{array}$$

By the definition of $\hat{\theta}$ in (4) and the Taylor expansion,

$$\begin{array}{rcl} 0 & \leq & l(\hat{\theta}) - l(\hat{\theta} \pm \varepsilon_N\Pi_N\nu^*) = l(\theta(0)) - l(\theta(1)) = -\left.\frac{\partial l(\theta(t))}{\partial t}\right|_{t=0} - \frac{1}{2}\left.\frac{\partial^2 l(\theta(t))}{\partial t^2}\right|_{t=s}, \text{ for some } s \in [0,1] \\ & = & \pm\varepsilon_N \frac{1}{N}\sum_{i=1}^N \frac{dl_i(\hat{\theta})}{d\theta'}\,[\Pi_N\nu^*] + \frac{1}{2}\frac{1}{N}\sum_{i=1}^N \frac{d^2\,l_i(\theta(s))}{d\theta'd\theta}\,[\pm\varepsilon_N\Pi_N\nu^*, \pm\varepsilon_N\Pi_N\nu^*] \\ & = & \pm\varepsilon_N \frac{1}{N}\sum_{i=1}^N \frac{dl_i(\hat{\theta})}{d\theta'}\,[\Pi_N\nu^*] + \frac{1}{2}E\,\frac{d^2\,l_i(\theta(s))}{d\theta'd\theta}\,[\pm\varepsilon_N\Pi_N\nu^*, \pm\varepsilon_N\Pi_N\nu^*] \\ & & + \frac{1}{2}\left\{\frac{1}{N}\sum_{i=1}^N \frac{d^2\,l_i(\theta(s))}{d\theta'd\theta}\,[\pm\varepsilon_N\Pi_N\nu^*, \pm\varepsilon_N\Pi_N\nu^*] - E\,\frac{d^2\,l_i(\theta(s))}{d\theta'd\theta}\,[\pm\varepsilon_N\Pi_N\nu^*, \pm\varepsilon_N\Pi_N\nu^*]\right\} \end{array}$$

We follow Chen et al. (2006) and show that

$$\frac{1}{N}\sum_{i=1}^N \frac{dl_i(\theta_o)}{d\theta'}\,[\Pi_N\nu^* - \nu^*] = o_p(N^{-1/2}) \tag{11}$$

and that, uniformly over $\theta(s)$ in a neighborhood of θ_o with $||\theta(s) - \theta_o|| = O(\delta_N)$,

$$E\,\frac{d^2\,l_i(\theta(s))}{d\theta'd\theta}\,[\pm\varepsilon_N\Pi_N\nu^*, \pm\varepsilon_N\Pi_N\nu^*] = \pm\varepsilon_N\langle\hat{\theta} - \theta_o, \nu^*\rangle \pm \varepsilon_N\, o_p(N^{-1/2}) \tag{12}$$

and

$$\frac{1}{N}\sum_{i=1}^N \frac{d^2\,l_i(\theta(s))}{d\theta'd\theta}\,[\pm\varepsilon_N\Pi_N\nu^*, \pm\varepsilon_N\Pi_N\nu^*] - E\,\frac{d^2\,l_i(\theta(s))}{d\theta'd\theta}\,[\pm\varepsilon_N\Pi_N\nu^*, \pm\varepsilon_N\Pi_N\nu^*] = \varepsilon_N\, o_p(N^{-1/2}) \tag{13}$$

It will then follow that

$$0 \leq l(\hat{\theta}) - l(\hat{\theta} \pm \varepsilon_N\Pi_N\nu^*) = \pm\varepsilon_N \frac{1}{N}\sum_{i=1}^N \frac{dl_i(\theta_o)}{d\theta'}\,[\nu^*] \pm \varepsilon_N\langle\hat{\theta} - \theta_o, \nu^*\rangle \pm \varepsilon_N\, o_p(N^{-1/2})$$

And, since $\varepsilon_N = o(N^{-1/2}) > 0$, we have

$$\begin{aligned}\sqrt{N}\langle\hat{\theta}-\theta_o,\nu^*\rangle &= \tfrac{1}{\sqrt{N}}\left(\sum_{i=1}^N \tfrac{dl_i(\theta_o)}{d\theta'}\,[\nu^*] - E\tfrac{dl_i(\theta_o)}{d\theta'}[\nu^*]\right)+o_P(1)\\ &\Rightarrow\ N(0,||\nu^*||^2),\end{aligned}$$

where $E\left(\frac{dl_i(\theta_o)}{d\theta'}[\nu^*]\right)=0$ and $||\nu^*||^2 = Var\left(\frac{dl_i(\theta_o)}{d\theta'}[\nu^*]\right)$. Now, since $\lambda'(\hat{\beta}-\beta_o)=\langle\hat{\theta}-\theta_o,\nu^*\rangle$, the conclusion of the theorem follows by the Cramér-Wold device. What remains is to show (11)-(13).

Equation (11) holds by Assumption A4(A), since $||\Pi_N\nu^*-\nu^*||=o(1)$. To show (12), note that, under Assumption A4(B), uniformly over $\theta(s)$ in a neighborhood of θ_o with $||\theta(s)-\theta_o||=O(\delta_N)$,

$$E\,\frac{d^2\,l_i(\theta(s))}{d\theta' d\theta}\,[\pm\varepsilon_N\Pi_N\nu^*,\pm\varepsilon_N\Pi_N\nu^*]\le\langle\hat{\theta}-\theta_o,\pm\varepsilon_N\Pi_N\nu^*\rangle=\pm\varepsilon_N\langle\hat{\theta}-\theta_o,\Pi_N\nu^*\rangle$$

But, by Assumption A4(A), $\langle\hat{\theta}-\theta_o,\Pi_N\nu^*-\nu^*\rangle=o_p(N^{-1/2})$. Thus, $\pm\varepsilon_N\langle\hat{\theta}-\theta_o,\Pi_N\nu^*\rangle=\pm\varepsilon_N\langle\hat{\theta}-\theta_o,\nu^*\rangle\pm\varepsilon_N o_p(N^{-1/2})$. For showing (13), recall that

$$\frac{d^2\,l_i(\theta(s))}{d\theta' d\theta}\,[\pm\varepsilon_N\Pi_N\nu^*,\pm\varepsilon_N\Pi_N\nu^*]=l_i(\hat{\theta}\pm\varepsilon_N\Pi_N\nu^*)-l_i(\hat{\theta})\pm\varepsilon_N\frac{dl_i(\hat{\theta})}{d\theta'}[\Pi_N\nu^*]$$

Now, for some $\theta(\tau)$ between $\hat{\theta}$ and $\hat{\theta}\pm\varepsilon\Pi_N\nu^*$, write $l_i(\hat{\theta}\pm\varepsilon_N\Pi_N\nu^*)-l_i(\hat{\theta})=\pm\varepsilon_N\frac{dl_i(\theta(\tau))}{d\theta'}[\Pi_N\nu^*]$. Then, the left hand side of (13) can be written as

$$\pm\varepsilon_N\left\{\frac{1}{N}\sum_{i=1}^N\frac{d\,l_i(\theta(\tau))}{d\theta'}[\Pi_N\nu^*]-E\frac{d\,l_i(\theta(\tau))}{d\theta'}[\Pi_N\nu^*]\right\}\pm\varepsilon_N\left\{\frac{1}{N}\sum_{i=1}^N\frac{d\,l_i(\hat{\theta})}{d\theta'}[\Pi_N\nu^*]-E\frac{d\,l_i(\hat{\theta})}{d\theta'}[\Pi_N\nu^*]\right\},$$

which is $\pm\varepsilon_N o_p(N^{-1/2})$.

Proof of Theorem 2: We apply the method of Severini and Tripathi (2001). To make it easier to follow for those who know their method, we use their notation and also specify our equivalents of their objects. For some $t_o>0$ let $\theta(t)$ denote a curve from $[0,t_o]$ into Θ such that $\theta(0)=\theta_o$. The curve we consider is $\theta(t)=\theta_o+t\nu$, for any $\nu\in V$. Let $\dot{\theta}$ denote the slope of $\theta(t)$ at $t=0$, i.e. $\dot{\theta}$ is tangent to the set Θ at θ_o. For our case, $\dot{\theta}=\nu$. Let $T(\Theta,\theta_o)$ denote the collection of all such tangents $\dot{\theta}'s$ and let $\bar{T}(\Theta,\theta_o)$ denote the linear closure of $T(\Theta,\theta_o)$, i.e. the tangent space. In our case, $\bar{T}(\Theta,\theta_o)=\bar{V}$.

The objective is to obtain the efficiency bound for estimating $\rho(\theta_o)=\lambda'\beta_o$. Stein (1956) is often credited for being first to suggest that the efficiency bound can be viewed as the upper bound on the asymptotic variance for estimating any one-dimensional subproblem of the original problem. Our one-dimensional subproblem is estimation of t, whose true value is zero. The score for estimating $t=0$ is $s_i=\left.\frac{dl_i(\theta_t)}{dt}\right|_{t=0}=\left.\frac{d\ln h(\mathbf{y}_i;\theta_t)}{dt}\right|_{t=0}=\frac{d\ln h(\mathbf{y}_i;\theta_o)}{d\theta}[\dot{\theta}]$. In our notation, this is just the directional derivative $\dot{l}(\theta_o)[\nu]$ for obsevation i, call it $\dot{l}_i(\theta_o)[\nu]$. Then, the Fisher information for estimating $t=0$ is given by $||\nu||^2=Es_i^2$.

We now look at those one-parameter subproblems that are informative about the feature of interest $\rho(\theta_o)$, specifically, we focus on those curves $\theta(t)$ that satisfy the restriction $\rho(\theta(t)) = t$. This means choosing among only those $\dot{\theta}'s$ that satisfy $\left.\frac{d\rho(\theta(t))}{dt}\right|_{t=0} = 1$, or equivalently, only those $\nu's$ for which $\dot{\rho}(\theta_o)[\nu] = 1$. A simplification that applies in our case is that $\dot{\rho}(\theta_o)[\nu] = \rho(\nu) = \lambda'\nu_\beta$. Then, for any consistent estimator $\hat{t}$, $AV\left\{\sqrt{N}\left[\rho(\theta(\hat{t})) - \rho(\theta_o)\right]\right\} = AV(\sqrt{N}\hat{t}) \geq ||\nu||^{-2}$. Now to obtain the semiparametric lower bound (SPLB) for estimating $\rho(\theta_o)$, we look for a ν that maximizes $||\nu||^{-2}$. As discussed in Severini and Tripathi (2001, p. 28), the maximization problem can be equivalently written as

$$\text{SPLB} = \sup_{\nu\in\bar{V}:\nu\neq 0,\lambda'\nu_\beta=1} ||\nu||^{-2} = \sup_{\nu\in\bar{V}:\nu\neq 0} \left|\left|\frac{\nu}{\lambda'\nu_\beta}\right|\right|^{-2} = \sup_{0\neq\nu\in\bar{V}} \frac{|\lambda'\nu_\beta|^2}{||\nu||^2} = \sup_{||\nu||=1} |\lambda'\nu_\beta|^2 = ||\dot{\rho}(\theta_o)[\nu]||_*^2,$$

where $||L(\nu)||_*$ is the norm of a continuous linear functional $L(\nu)$ on the tangent space.

Calculating the norm is usually easier by appealing to the Riesz representation theorem as done in the main text. Basically, instead we look for the representer of the functional. The Riesz representation theorem says that $||\dot{\rho}(\theta_o)[\nu]||_* = ||\nu^*||$, where ν^* as defined in (8). Thus, $\text{SPLB} = ||\nu^*||^2$.