Efficient estimation of parameters in marginals in semiparametric multivariate models*

Ivan Medovikov† Valentyn Panchenko‡ Artem Prokhorov§

January 2023

Abstract

We consider a general multivariate model where univariate marginal distributions are known up to a parameter vector and we are interested in estimating that parameter vector without specifying 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 QMLE estimator. If we assume a parametric copula (other than independence) we obtain a full MLE, which is efficient but only under a correct copula specification and may be biased if the copula is misspecified. Instead we propose a sieve MLE estimator (SMLE) which improves over QMLE but does not have the drawbacks of 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 relevant 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. We demonstrate practical value of the new estimator with several applications. First, we apply SMLE in an insurance context where we build a flexible semi-parametric claim loss model for a scenario where one of the variables is censored. As in simulations, the use of SMLE leads to tighter parameter estimates. Next, we consider financial risk management examples and show how the use of SMLE leads to superior Value-at-Risk predictions. The paper comes with an online archive which contains all codes and datasets.

*Helpful comments of seminar participants at University of Toronto, University of Pittsburgh, UNSW Sydney, Concordia University, QMF, International Panel Data and FESAMES are gratefully acknowledged.

†Brock University, St Catharines ON L2S 3A1, Canada; email: imedovikov@brocku.ca

‡Economics, UNSW Business School, Sydney NSW 2052, Australia; email: valentyn.panchenko@unsw.edu.au

§University of Sydney Business School & CEBA & CIREQ; Sydney NSW 2006, Australia; email: artem.prokhorov@sydney.edu.au

1 Introduction

Consider an m-variate random variable Y with joint pdf $h(y_1, \ldots, y_m)$. Let $f_1(y_1; \beta_1), \ldots, f_m(y_m; \beta_m)$ denote the corresponding marginal pdf's, known up to a parameter vector. 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 several cross-sections, each of which represents a different but not unrelated random variable. A classic copula application is the joint modelling of insurance claim payments and claim-related expenses (see e.g., Frees and Valdez, 1998). There is a well specified marginal for each cross section, e.g., a well justified family of distributions of historical insurance losses, 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).

As an alternative empirical setting, consider multivariate financial applications, where interest is in capturing the temporal dependence between processes (see, e.g., Chen and Fan, 2006a,b; Hafner and Reznikova, 2010). In such a setting, it is of much practical importance to obtain improved estimates of a feature of an univariate conditional distribution such as Value-at-Risk, by accounting for dependence of one time series with others (see, e.g., Pitt and Walker, 2005, for several such applications in state-space modelling).

The 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 some financial applications where it is important to parameterize dependence. In our setting, dependence is used solely to provide more precision in estimating marginal parameters, so we study the converse problem.

Our starting point is the marginals known up to a parameter vector. However this does not preclude some misspecification in the marginals. In particular, our results still hold in cases when misspecification does not lead to inconsistency of estimation for the feature of interest. Because we deal with generic likelihoods, in essence we allow for the marginals to be incorrect as long as they have zero-mean score functions at the pseudo-true parameter (White, 1982).

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

$$\ln h(y_1, \ldots, y_m; \beta) = \sum_{j=1}^{m} \ln f_j(y_j; \beta_j) + \ln c(F_1(y_1; \beta_1), \ldots, F_m(y_m; \beta_m)), \tag{1}$$

where $c(\cdot)$ is a copula density, $F_j(\cdot)$ denotes the corresponding marginal cdf and where we collect all parameters of the marginals in one vector β but allow for each marginal to depend on distinct subvectors of β. 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. This is valid for any m but in simulations and applications we focus on the values $m = \{2, 3\}$ to keep the nonparametric task of estimating the copula component manageable.

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 one. However, QMLE ignores the dependence information and is not efficient if marginals are not independent. For highly dependent marginals, the efficiency loss relative to the correctly specified full likelihood 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. We note that the marginals do not have to have common parameters for this result to apply.[1]

There are numerous estimators that improve on QMLE but remain robust to dependence by not specifying it. For example, Prokhorov and Schmidt (2009) propose stacking the score functions from the marginal distributions and applying the Generalized Method of Moments (GMM) machinery to achieve the improvements via the use of correlation between the marginal scores[2]; Nikoloulopoulos et al. (2011) use a similar approach and construct a weighted sum of the marginal scores by fitting and discretizing a multivariate normal model for the scores. These estimators are simple because they are based on linear combinations of the marginal scores. They are somewhat restrictive in that they cannot attain full efficiency unless the use of the true copula cannot improve upon the use of a linear combination of marginal scores. We return to this point later.

[1]Under dependence, random variables are constrained by an additional functional relationship that typically helps to identify the parameters of the marginals. Specifically, in the bivariate case, $y_1 = F_1^{-1}(C^{-1}(\eta|F_2(y_2;\beta_2));\beta_1)$, where $C^{-1}(\eta|F_2(y_2;\beta_2))$ is the generalized inverse of conditional copula $C(u_1|u_2) = \frac{\partial C(u_1,u_2)}{\partial u_2}$ and $\eta \in [0,1]$ is a uniform random variable. See some examples with specific marginals for the simpler cases of extreme dependence in Section 3.

[2]The GMM is also known as the method of estimating equations (see, e.g., Hansen, 1982; Godambe and Thompson, 1978)

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. Hao et al. (2018) proposed a test of copula redundancy. Essentially, a parametric copula is redundant for the 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 consequence, significant efficiency gains due to the nonlinearity of the copula score in terms of the marginal scores 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), including bounds for rank-based copula estimators (see, e.g., Segers et al., 2014; Hoff et al., 2014). The paper is similar to Hu et al. (2017) in that they also use a sieve MLE involving the Bernstein polynomial and discuss convergence and efficiency.[3] However, they work with copula functions, not copula densities, which complicates monotonicity restrictions, and their setup is restricted to proportional hazard models; they do not discuss relative efficiency of SMLE, do not go beyond two dimensions or derive the Riesz representer.

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 actuarial application in two dimensions and a financial application in two and three dimensions. 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 $\Gamma = \{c(\mathbf{u}) : [0,1]^m \to [0,1], \int_{[0,1]^m} c(\mathbf{u})d\mathbf{u} = 1, \int_{[0,1]^{m-1}} c(\mathbf{u})d\mathbf{u}_{-\ell} = 1, \forall \ell\}$, where

[3] We thank an anonymous referee for pointing out the existence of this paper to us.

$\mathbf{u}_{-\ell}$ 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. See Appendix A for the basics of sieve MLE.

Let Γ_N denote a sequence of approximating spaces, called sieves, such that $\bigcup_N \Gamma_N$ is dense in Γ. One of the challenges of SMLE in our setting is ensuring that Γ_N consists of proper copula pdfs, that is, non-negative functions that integrate to one and have uniform marginals. Exponential or quadratic transformations are often used 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. For example, Anderson et al. (2021) show that very few of the popular nonparametric copula estimators satisfy this property in finite samples. 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. A sieve which does not require any transformation to satisfy the proper copula conditions and has meaningful parameters is the Bernstein-Kantorovich polynomial (see, e.g., Sancetta and Satchell, 2004).

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).[4] In order to ensure that $c_{J_N}(\mathbf{u})$ is a copula density, i.e. that its marginals are uniform, we further require that $\sum_{\mathbf{v}_{-\ell}|v_\ell} \omega_{\mathbf{v}} = 1/J_N$, where multiple summations are performed over all elements of $\mathbf{v}$ except v_ℓ, $\ell = 1, \ldots, m$ for each fixed value of v_ℓ, where $v_\ell = 0, \ldots J_N - 1$. Hence, there are $J_N \times m$ of these restrictions in total.

The weights $\omega_{\mathbf{v}}$ are akin to 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

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

Then, the Bernstein-Kantorovich polynomial sieve can be viewed as a smoothed copula histogram where smoothing is done by the product of beta-densities. Alternatively, it can be

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

viewed as a mixture of a product of beta-densities in u (see, e.g., Burda and Prokhorov, 2014).

Sancetta (2007) derives the rates of convergence of the Bernstein-Kantorovich copula to the true copula. Hirukawa et al. (2020) explore weak and strong uniform convergence of beta kernels on expanding compact sets on $(0, 1)$. Petrone and Wasserman (2002) and Burda and Prokhorov (2014) establish consistency of the Bernstein-Kantorovich polynomial when used as a prior on the space of densities on $[0, 1]^m$ in a univariate and multivariate 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 known to converge to the probability limit of the empirical copula density 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).[5]

This sieve is particularly attractive in our setting because of the uniform rate of convergence results available for c_{J_N} and because of the empirical copula density 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. Other potential explanations for the good performance in economics, finance, actuarial science and

[5]In practice, the choice of J_N is important to the extent to which it affects the bias-variance trade-off in finite samples: as shown by Sancetta and Satchell (2004), the Bernstein-Kantorovich sieve has bias of the order $O(J_N^{-1})$, which is the same as for a histogram or kernel-smoothers, but variance of order $O(J_N^{m/2})$ inside the hypercube, which is a square-root of the rate for a histogram or kernel estimator. The theoretically optimal order for J_N in the MSE sense is $O(N^{\frac{2}{m+4}})$, which is greater than for standard nonparametric estimators such a histogram or first-order kernels, implying relatively little smoothing required for this sieve. Suboptimal growth of J_N affects the balance of bias and variance but has no effect on semiparametric efficiency of $\hat{\beta}$ as long as $J_N \to \infty$, $\frac{J_N}{N} \to 0$ and Assumption A4 holds.

risk management are that such data have inhomogeneous dependence structures, are not highly correlated and sparse (see, e.g., Diers et al., 2012).

2.2 Asymptotic Properties

Let the sieve for $\Theta = B \times \Gamma$ be denoted by $\Theta_N = B \times \Gamma_N$, where Γ_N contains a generic vector of copula parameters γ, and let $\theta = (\beta', \gamma)$. For the special case of the Bernstein-Kantorovich copula, $\gamma = \omega_{\mathbf{v}}$.

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). In the discussion of these assumptions we focus on what is new to our copula-based settings.

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) $\Gamma = \{c = \exp(g) : g \in \Lambda^r([a,b]^m), \int c(\mathbf{u})d\mathbf{u} = 1, \int_{[0,1]^{m-1}} c(\mathbf{u})d\mathbf{u}_{-\ell} = 1, \forall \ell\}$, where $\Lambda^r([a,b]^m)$ denotes the Hölder class of r-smooth functions on $[a,b]^m$, $\forall [a,b] \subset (0,1)$, $r > 1/2$, 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 inside the hypercube

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

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. We exclude the boundaries of the hypercube which means we exclude from consideration the edges of $[0,1]^m$ where copula densities can be unbounded.[6] Commonly used densities, including copula densities, belong to the Hölder class on $[a,b]^m$, and various linear sieves, as well as the Bernstein-Kantorovich polynomial sieve, are known to approximate such functions well. Commonly used copulas satisfy the stronger property of Lipschitz continuity (see, e.g., Siburg and Stoimenov, 2008) but this property does not translate to copula densities. $\Lambda^r([a,b]^m)$ is one of the most popular function classes in nonparametric estimation literature (see, e.g., Horowitz, 1998; Chen, 2007).

In our semi-parametric settings, the initial parameter 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.

[6] An alternative is to employ at the edges an expanding set sequence (see, e.g., Hirukawa et al., 2020) or a trimming or weighting scheme (see, e.g., Hirukawa et al., 2020; Hill and Prokhorov, 2016). We leave such approaches for future work.

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,

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

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, $\langle \nu_1, \nu_2 \rangle = E\left[\dot{l}(\theta_o)[\nu_1]\dot{l}(\theta_o)[\nu_2]\right]$ and $||\nu|| = \sqrt{\langle \nu, \nu \rangle}$, 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 matrix for β. For our specific setting, this minimization problem can be written as follows:

$$\inf_{g_q} E\left[\sum_{j=1}^{m}\left\{\frac{\partial \ln f_j(y_j,\beta_o)}{\partial\beta_q} + \left(\frac{1}{c(\mathbf{u})}\frac{\partial c(\mathbf{u})}{\partial u_j}\right)\Big|_{u_k=F_k(y_k,\beta_o)} \frac{\partial F_j(y_j,\beta_o)}{\partial\beta_q}\right\} + \left(\frac{1}{c(\mathbf{u})}g_q(\mathbf{u})\right)\Big|_{u_k=F_k(y_k,\beta_o)}\right]^2, \tag{4}$$

where $E\left[\frac{1}{c(\mathbf{u})}g_q(\mathbf{u})\right] = 0$. Let g_q^* denote the solution of (4), $q = 1,\ldots,p$, and let $g^* = (g_1^*,\ldots,g_p^*)$.

We can now find the sup by writing

$$\begin{aligned}\sup_{\nu\neq 0,\nu\in\bar{V}} \frac{|\lambda'\nu_\beta|^2}{||\nu||^2} &= \sup_{\nu\neq 0,\nu\in\bar{V}}\left\{|\lambda'\nu_\beta|^2\left(E\left[\dot{l}(\theta_o)[\nu]^2\right]\right)^{-1}\right\} \\ &= \lambda'\left(ES_\beta S_\beta'\right)^{-1}\lambda,\end{aligned} \tag{5}$$

where

$$\begin{aligned} S_\beta' &= \sum_{j=1}^m\left\{\frac{\partial \ln f_j(y_j,\beta_o)}{\partial\beta'} + \left(\frac{1}{c(\mathbf{u})}\frac{\partial c(\mathbf{u})}{\partial u_j}\right)\Big|_{u_k=F_k(y_k,\beta_o)}\frac{\partial F_j(y_j,\beta_o)}{\partial\beta'}\right\} \\ &\quad + \left(\frac{1}{c(\mathbf{u})}g^*(\mathbf{u})\right)\Big|_{u_k=F_k(y_k,\beta_o)} \\ g^* &= (g_1^*,\ldots,g_p^*) \quad \text{and} \quad E\left[\frac{1}{c(\mathbf{u})}g_q^*(\mathbf{u})\right] = 0. \end{aligned} \tag{6}$$

Note that the second equality in (5) is true because g^* is the minimizer of $E\left[\dot{l}(\theta_o)[\nu]^2\right]$ over ν_γ at the true β_0. So $\rho(\theta) = \lambda'\beta$ is bounded if and only if $ES_\beta S_\beta'$ in (5) is a finite and positive

definite matrix.

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

It is worth returning to the parametric setting to illustrate the intuition behind Assumption A3. In essence, $ES_\beta S_\beta'$ is the marginal Fisher information and Assumption A3 can be viewed as a non-redundancy condition of the true copula for the estimation of β (see Prokhorov and Schmidt, 2009, Section 4). Aside from technical failures such as moment non-existence, it assumes away cases when knowledge of the true copula cannot improve precision in the MLE of β in principle. Prokhorov and Schmidt (2009) show that these cases are problem specific and rare – this happens only if the copula score happens to be a linear combination of the marginal scores of β. For example, for a bivariate normal with a common mean and known correlation, the copula score for the mean is a linear combination of the marginal scores for the mean. However, this is not the case if the normal copula is replaced by the FGM copula or any other commonly used copula function with known dependence parameter (see, Prokhorov and Schmidt, 2009, Examples 1, 5 and 6).

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 for our settings.

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

$$\nu^* = \left(I, g^{*'} \right)' \left(E S_\beta S_\beta' \right)^{-1} \lambda \tag{7}$$

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

follows

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

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 $(\delta_N)^w = o(N^{-1/2})$, $w > 1$; (B) there exists $\Pi_N\nu^* \in V_N - \{\theta_o\}$ such that $\delta_N||\Pi_N\nu^* - \nu^*|| = o(N^{-1/2})$ and (C) that, for any $\theta : ||\theta-\theta_o|| = O_p(\delta_N)$, the additional conditions on the second-order derivatives stated in the Appendix hold.

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**(A) covers all commonly encountered sieves. For example, for the trigonometric sieve, Shen and Wong (1994) show that its order of convergence is $O_p(N^{-r/(2r+1)})$, where r is the Hölder exponent; for Bernstein-

Kantorovich polynomial sieves, Sancetta and Satchell (2004) show that its rate of convergence is $O_p(N^{-4/(m+4)})$ within the hypercube, where m is the dimension; Bouezmarni et al. (2010) extend the results of Sancetta and Satchell (2004) to α-mixing data. Assumptions **A4**(B)-(C) are technical assumptions that control smoothness of the Riesz representer and the second order term in the expansion of the log-likelihood (see Chen and Fan, 2006b, for details).

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

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

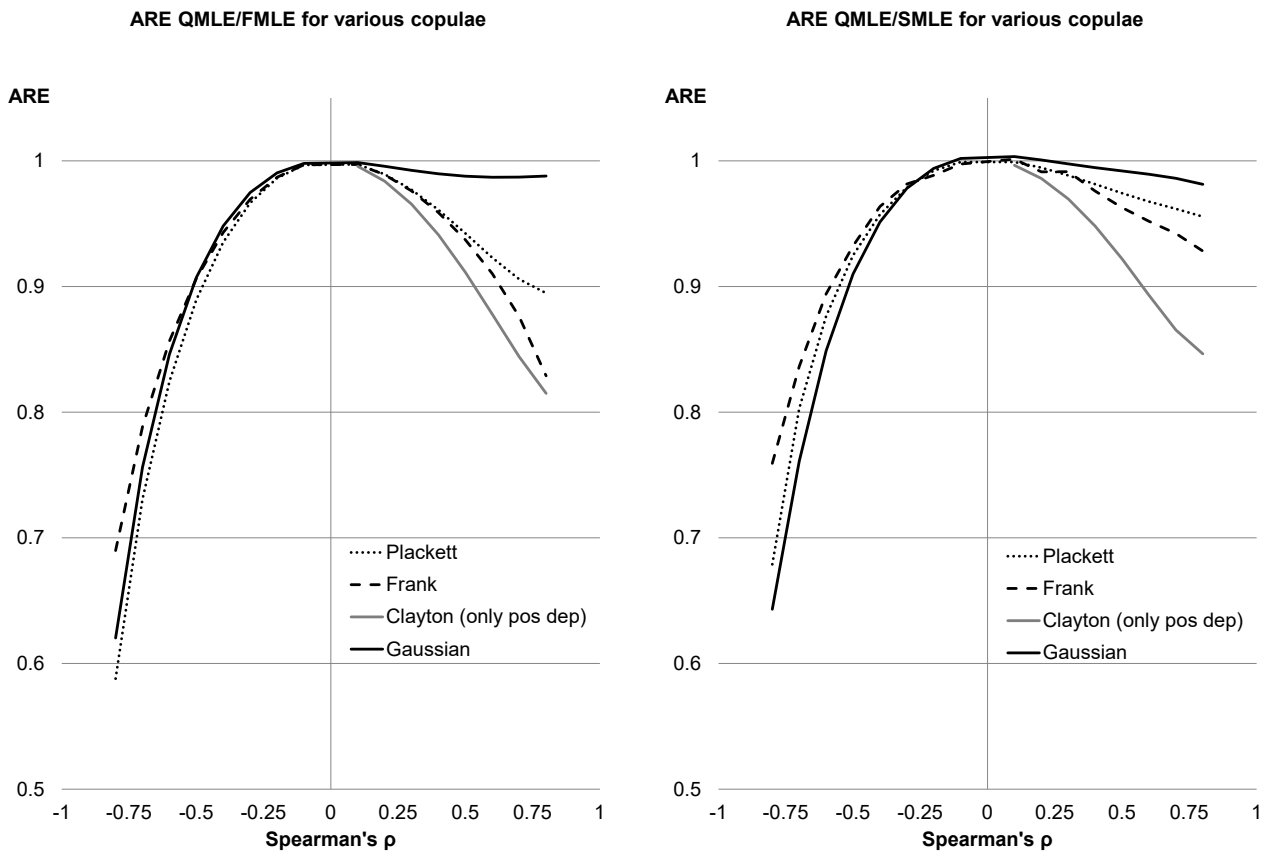
Proof. See Appendix B for all proofs.

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.[7] In our settings, we obtain consistent estimators g_q^* as solutions to the following problem

$$\arg \min_{g_q \in \mathbf{A}_N} \sum_{i=1}^{N} \left[\sum_{j=1}^{m} \left\{ \frac{\partial \ln f_j(y_{ji}, \hat{\beta})}{\partial \beta_q} + \left(\frac{1}{\hat{c}(\hat{\mathbf{u}}_i)} \frac{\partial \hat{c}(\hat{\mathbf{u}}_i)}{\partial u_j} \right) \Big|_{\hat{u}_{ki} = F_k(y_{ki}, \hat{\beta})} \frac{\partial F_j(y_{ji}, \hat{\beta})}{\partial \beta_q} \right\} + \frac{1}{\hat{c}(\hat{\mathbf{u}}_i)} g_q(\hat{\mathbf{u}}_i) \Big|_{\hat{u}_{ki} = F_k(y_{ki}, \hat{\beta})} \right]^2, \quad q = 1, \ldots, p \tag{8}$$

[7] An alternative estimator of $E[S_\beta S_\beta']^{-1}$ was proposed by Ackerberg et al. (2012, 2014). It uses the covariance matrix of all $p + J_N^m$ model parameters (both parameters in the marginal and in the copula). The upper left $p \times p$ block of its inverse is used a variance estimator. 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.



(a) (b)

Figure 1: Asymptotic relative efficiency of QMLE to FMLE (a) and SMLE (b).

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 $\int g_q(\mathbf{u})/\hat{c}(\mathbf{u})\, d\mathbf{u} = 0$.

Specific aspects of our implementation of the sieve estimator are discussed in Appendix C.

3 Simulations

We focus on the two-dimensional distributions first and then extend the simulations to a three-dimensional case.[8] Our simulation study is inspired by Joe (2005) who studies the

[8] A Matlab code implementing the Bernstein-Kantorovich sieve and other codes used in this and the next sections are available at `http://research.economics.unsw.edu.au/vpanchenko/software/scopula.zip`.

asymptotic relative efficiency (ARE) of copula based MLE, i.e. the ratio of the asymptotic variance of FMLE to that of QMLE. He shows that ARE depends on the specification of marginals and copula as well as on the strength of dependence. Moreover, for asymmetric marginal distributions, e.g., exponential, he finds that ARE for strongly negatively dependent data is much larger than for strongly positively dependent data holding the same absolute dependence strength.[9] We start our simulations from the bivariate DGPs with exponential marginals with distinct means $\mu_1 = 0.5, \mu_2 = 1$;[10] the dependence is modeled with various commonly used bi-variate copulae, i.e., Gaussian, Clayton, Plackett, and Frank. Figure 1 reports AREs – panel (a) QMLE/FMLE and panel (b) QMLE/SMLE – as a function of dependence strength measured by Spearman's ρ for the various copulae we use. Spearman's ρ varies in the range from -0.8 to 0.8. Note that we use the Clayton copula only for positive dependence.[11] The SMLE asymptotic variance is estimated using (8) for a sample of 1,000,000 observations, where we use the tensor product sieve with cosine basis functions without the constant term to approximate g_q. The number of sieve elements is $10 \times 10 = 100$.

Figure 1 confirms that there is a scope for improvement over the QMLE and that the largest gains are in the case of strong negative dependence. Naturally, the efficiency gains reported in Figure 1 using FMLE (panel a) are higher than those obtained using SMLE (panel b). As expected the AREs of both FMLE and SMLE are near one (subject to some estimation noise)

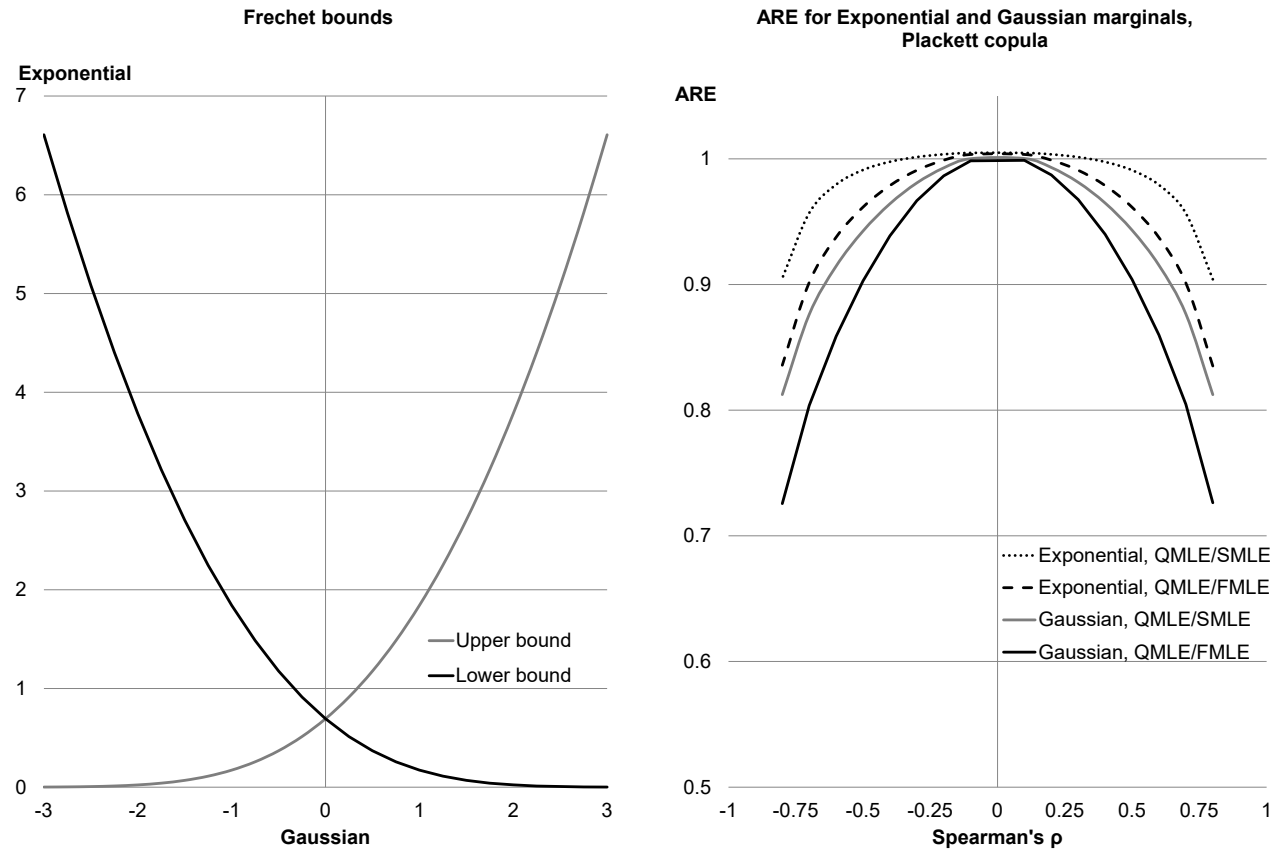
[9]Introducing a negative dependence in the DGP with two exponential marginals makes them skewed in the opposite directions. Accounting for the dependence in this case substantially helps with estimating the parameters of the marginals. We illustrate this later using the Fréchet bounds.

[10]Note that it is easy to show analytically that for a multivariate distribution with exponential marginals and an arbitrary copula function, ARE of QMLE relative to FMLE does not depend on the parameters in the marginals. Our simulations suggest that the same holds for the SMLE. For generic marginals, ARE depends on both parameters in the marginals and the dependence parameter.

[11]The Clayton copula can be extended to incorporate negative dependence, but certain regions would have zero density (Joe, 1997, p. 158).

in the case of independence, when we expect no gains over QMLE. 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, SMLE asymptotic variance bounds with copula parameters corresponding to Spearman's $\rho = -0.8$ suggest improvements of 31-54% over QMLE asymptotic variance depending on the copula. In the case of strong positive dependence FMLE does not show much efficiency gain over QMLE, which also agrees with Joe (2005). The simulations summarized in Figure 1 panel (b) show that similar patterns hold for SMLE. Joe (2005, Section 3) provides a detailed, identification-based, explanation for the asymmetry in ARE of FMLE with respect to the sign of ρ by considering limiting dependence cases known as upper and lower Fréchet bounds. At a bound, there is an exact functional relation (different for the upper and lower bound) between the two dependent variables, $y_1 = h(y_2; \beta_1, \beta_2)$. If parameters of the marginals β_1 and β_2 can be identified from this functional relationship, efficiency gains can be expected and this happens for some asymmetric distribution families in the Fréchet lower bound, and not for others.

If any of the marginals is symmetric, the shape of the Fréchet upper and lower bounds will be the same (reflected around the point of symmetry) and there will no difference in efficiency for the negative and positive dependencies of the same scale. We illustrate this by simulations with the bivariate DGP that has an exponential and Gaussian marginal. Figure 2, panel (a), shows the Fréchet upper and lower bounds for the bivariate distribution that has the exponential marginal with mean 1 and the Gaussian marginal with mean 0 and variance 1. Panel (b) shows the AREs of QMLE to FMLE and QMLE to SMLE for this DGP. For



(a) (b)

Figure 2: Fréchet bounds (a) and ARE of QMLE to FMLE and SMLE (b), symmetric case.

simplicity, the variance parameter is assumed to be known for the Gaussian marginal.

We ran a similar analysis for other marginal distributions, such as Gaussian with both mean and variance as unknown parameters, skew-Gaussian, Pareto, logistic, gamma and some of their combinations (not reported for brevity). The general patters of the efficiency gains are similar to those reported earlier.

Next we investigate in detail the performance of SMLE, FMLE and QMLE for a fixed value of Spearman's ρ. We use the exponential marginals and the Plackett copula as the DGP and set the true parameter values in the marginals at $\mu_1 = 0.5$ and $\mu_2 = 1$ and in

	$\mu_1 = 0.5$			$\mu_2 = 1$		
	FMLE	SMLE	QMLE	FMLE	SMLE	QMLE
Mean	0.5004	0.4987	0.5001	0.9992	0.9952	0.9991
Var $\times N$	0.1724	0.1917	0.2635	0.6956	0.7506	1.0346
MSE $\times N$	0.1725	0.1935	0.2635	0.6962	0.7735	1.0354
AVar $\times N$	0.1582	0.1797	0.2500	0.6329	0.7193	1.0000

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

the copula at $\gamma = 0.05$. This illustrates the case when marginals are not restricted to have identical parameters and implies moderate negative dependence with Spearman's ρ of -0.77. The sample size is $N = 1,000$ and the number of simulations is $1,000$.

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, scaled by N. Under Var, we report sample variance estimates while under AVar, we report estimates of the asymptotic variance obtained using a solution to (8). The number of elements in the Bernstein-Kantorovich sieve in one dimension is $J_N = 10$ and in total $10 \times 10 = 100$. This number minimizes the sum of mean-squared errors for both estimates (see Table 2). A key feature of the table is that SMLE shows substantial improvement over QMLE. The sample variance is close to the asymptotic variance bound.

One of the practical problems we face in implementing SMLE is the choice of the degree of polynomial J_N. 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 AIC and BIC, and we compare these criteria.

In particular, we investigate how the SMLE estimates change with the number of sieve

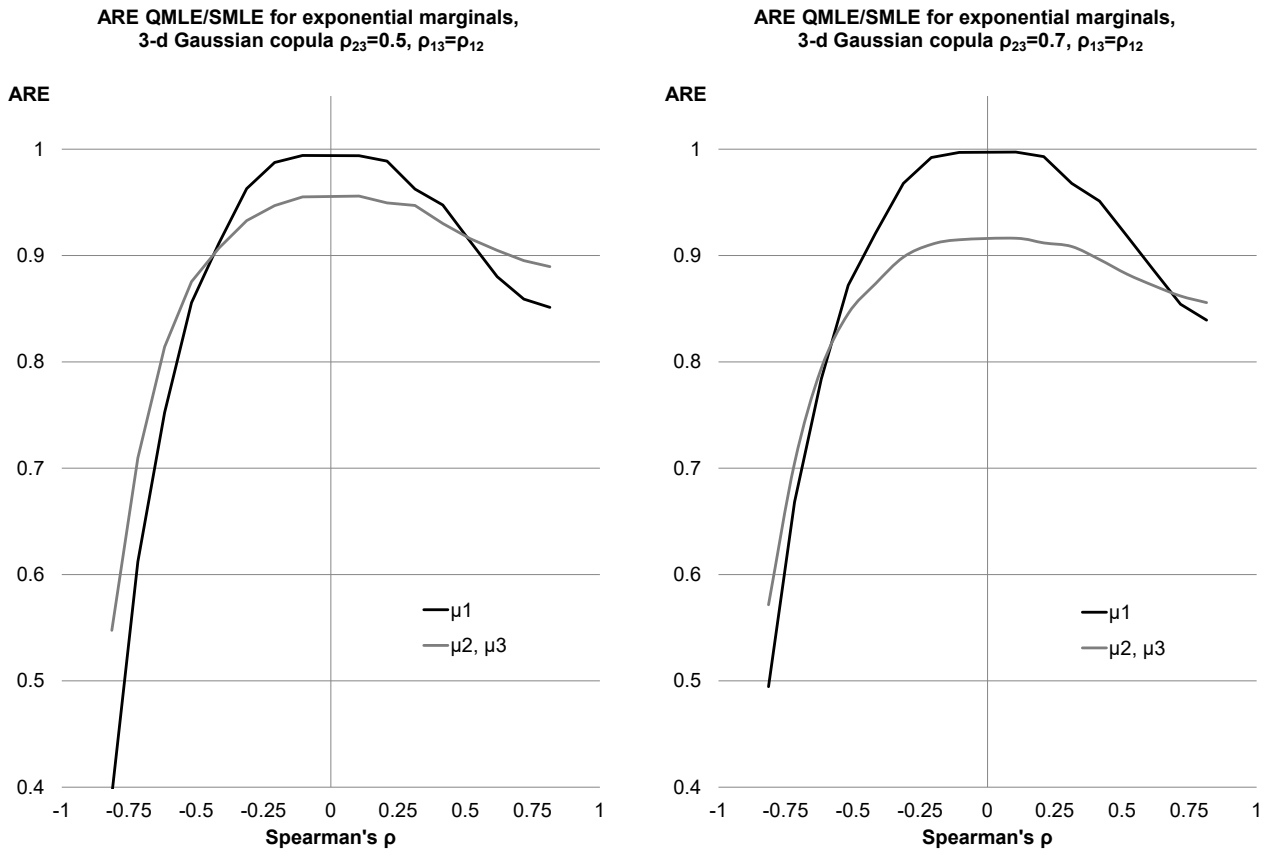
J_N	Mean1	Mean2	Var1	Var2	MSE1	MSE2	sumMSE	LogL	AIC	BIC	run-time
2	0.4906	0.9798	0.2334	0.9179	0.3209	1.3245	1.6454	-1106.44	2218.87	2233.60	0.008
3	0.4916	0.9817	0.2213	0.8729	0.2925	1.2092	1.5017	-1007.75	2027.51	2056.95	0.007
4	0.4938	0.9861	0.2116	0.8548	0.2499	1.0482	1.2981	-948.42	1918.84	1972.83	0.013
5	0.4956	0.9896	0.2062	0.8338	0.2251	0.9417	1.1667	-909.34	1854.67	1943.01	0.030
6	0.4968	0.9915	0.2012	0.8229	0.2115	0.8946	1.1062	-881.48	1816.96	1949.47	0.078
7	0.4976	0.9931	0.1944	0.8020	0.2000	0.8502	1.0501	-860.89	1797.78	1984.27	0.202
8	0.4982	0.9940	0.1919	0.7775	0.1953	0.8134	1.0087	-845.31	1792.62	2042.92	0.453
9	0.4985	0.9947	0.1901	0.7593	0.1924	0.7871	0.9795	-828.05	1788.09	2112.01	0.752
10	0.4987	0.9952	0.1917	0.7506	0.1935	0.7735	0.9669	-818.75	1803.49	2210.84	1.301
11	0.4989	0.9953	0.1914	0.7596	0.1927	0.7812	0.9739	-811.50	1827.00	2327.59	2.079
12	0.4989	0.9952	0.1899	0.7631	0.1911	0.7857	0.9768	-805.62	1857.23	2460.88	3.230
13	0.4985	0.9954	0.1945	0.7704	0.1968	0.7918	0.9886	-800.59	1893.18	2609.72	5.088
14	0.4985	0.9950	0.1979	0.7842	0.2002	0.8088	1.0089	-796.33	1934.66	2773.89	7.238
15	0.4983	0.9948	0.2014	0.8009	0.2044	0.8281	1.0326	-792.57	1981.14	2952.87	11.630

Table 2: Optimal number of sieve elements in SMLE

elements. Table 2 reports means, variances and MSEs, scaled by N, for the two estimates as well as the value of log-likelihood and popular model selection criteria, AIC and BIC. We also report average run-time in seconds for a specific J_N per one sample of 1,000 observations.[12] The value of log-likelihood, LogL, increases as sieve complexity grows, as expected. On average, BIC selects an under-parameterized model ($J_N = 5$), whereas AIC selects $J_N = 9$, which is close to $J_N = 10$ under which the smallest sum of MSEs is reached in the simulations. We also investigated K-fold cross-validation, but it was computationally expensive (45 minutes for $K = 10$) and did not provide any extra insights in addition to AIC. We also note a degree of stability of SMLE regardless of the tuning parameter, also noted by Sancetta and Satchell (2004, Table 2).

It is well known that nonparametric and semiparametric models are subject to the so-called "curse of dimensionality". The semiparametric copula sieve models are also affected by this issue. This also manifests in the non-linear relation between the run-time and the

[12]The simulations were performed on the UNSW computational cluster using one 16 CPU-cores node and Matlab parallel computing.



(a) (b)

Figure 3: ARE of SMLE for three dimensional copula.

number of sieve elements J_N (see the last column of Table 2). Nonetheless, we demonstrate that there are SMLE efficiency gains possible in three dimensions with simulations below and illustrate this point further with an application at the end of Section 4.

For easier comparison with the bivariate case we continue working with the exponential marginals labeled as $1, 2, 3$ with parameters $\mu_1 = 0.1$, $\mu_2 = 0.5$, and $\mu_3 = 1$, respectively. The dependence is specified with the Gaussian copula with three parameters, $\rho_{12}, \rho_{13}, \rho_{23}$, that reflect the dependence between the pairs of the marginals indicated by the subscripts. We vary ρ_{12} and ρ_{13} keeping them equal to each other and taking values from strong negative

dependence to strong positive dependence on the scale analogous to Figures 1 and 2, for comparison; the value of ρ_{23} is fixed. Figure 3 compares ARE of SMLE relative to QMLE for two scenarios: panel (a) $\rho_{23} = 0.5$ and panel (b) $\rho_{23} = 0.7$.[13] We observe that substantial efficiency gains can be realised in the trivariate case. AREs for the μ_1 on both panels are similar because these are mainly driven by $\rho_{12} = \rho_{13}$. AREs for μ_2 and μ_3 are essentially the same because of the assumed dependence structure. Higher ρ_{23} (panel b vs panel a) leads to higher efficiency gains (lower AREs) for the mid-range of $\rho_{12} = \rho_{13}$ but not on the extremes. As expected, FMLE reaches a somewhat higher efficiency in comparison to SMLE (not reported for brevity).

4 Empirical Examples

4.1 A Model For Insurance Claims

First, 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 1 if a given claim has surpassed the policy limit and 0 if not. For details of this classic data set, see Frees and Valdez (1998).

[13] There is a restriction on (negative) dependence range in the trivariate case, i.e., positive-definiteness of the corresponding correlation matrix. This motivates our choice for ρ_{23} given the full range we consider for $\rho_{12} = \rho_{13}$. Since the marginals are exponential, as in the bivariate case, AREs do not depend on the parameter of the marginals.

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:

$$F_j(Y_j) = 1 - \left(\frac{\lambda_j + Y_j}{\lambda_j}\right)^{-\theta_j}, \quad j = 1, 2. \tag{9}$$

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 note 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 (4) and a consistent estimate of the SMLE variance, $\hat{V}$, will now be

$$\arg \min_{g_q \in \mathbf{A}_N} \left[\sum_{i=1}^{N} (1-d_i) \left\{ \sum_{j=1}^{2} \left(\frac{\partial \ln f_j(y_{ji}, \hat{\beta})}{\partial \beta_q} + \frac{1}{\hat{c}(\hat{\mathbf{u}}_i)} \frac{\partial \hat{c}(\hat{\mathbf{u}}_i)}{\partial u_j} \frac{\partial F_j(y_{ji}, \hat{\beta})}{\partial \beta_q} \right) + \frac{1}{\hat{c}(\hat{\mathbf{u}}_i)} g_q(\hat{\mathbf{u}}_i) \right\} \right.$$
$$\left. + \sum_{i=1}^{N} d_i \left\{ \frac{\partial \ln f_2(y_{2i}, \hat{\beta})}{\partial \beta_q} - \frac{1}{1 - \hat{C}_2(\hat{u}_{1i}, \hat{u}_{2i})} \left(\sum_{j=1}^{2} \frac{\partial \hat{C}_2(\hat{\mathbf{u}}_i)}{\partial u_j} \frac{\partial F_j(y_{ji}, \hat{\beta})}{\partial \beta_q} + \int_0^1 g_q(s, \hat{u}_{2i})\, ds \right) \right\} \right]^2, \quad (10)$$

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 evaluate both g_q and its integral over u_1.

The estimates based on QMLE, FMLE with different copulas and SMLE, and their standard errors are given in Table 3. The FMLE estimator is based on a fully specified parametric joint likelihood including a copula. We estimate and compare the one-parameter Frank and Gumbel-Hougaard (G-H) copulas with dependence parameter γ as in Frees and Valdez (1998). Hua (2017) proposed a new two-parameter bivariate copula, called GGEE, which has both upper and lower tail dependence and showed that it was one of the best-performing copu-

	QML	Frank FML	G-H FML	Galambos FML	GGEE FML	SML
λ_1	14,443.05	14,562.05	14,040.84	14,227.91	14,653.92	14,367.29
	(1,515.08)	(1,498.47)	(1,351.17)	(1,372.70)	(1,504.56)	(1,480.17)
θ_1	1.135	1.115	1.122	1.131	1.152	1.117
	(0.076)	(0.073)	(0.067)	(0.068)	(0.075)	(0.072)
λ_2	15,133.34	16,708.36	14,223.69	13,973.37	15,250.40	15,444.65
	(1,744.97)	(1,900.45)	(1,405.77)	(1,359.51)	(1,697.91)	(1,726.95)
θ_2	2.223	2.312	2.119	2.094	2.226	2.240
	(0.183)	(0.190)	(0.143)	(0.138)	(0.176)	(0.177)
γ		3.158	1.453	0.727	1.848	
		(0.171)	(0.035)	(0.036)	(0.398)	
δ					0.807	
					(0.068)	
LogL	-31,951	-31,778	-31,749	-31,749	-31,750	-31,749

Table 3: QMLE, FMLE, and SMLE for insurance claims and related expenses. Standard errors are reported in parentheses.

las in a comprehensive comparison of various copulas for the same dataset. We include the GGEE copula along with the one-parameter Galambos copula which also exhibited a good fit for these data. The QMLE and FMLE standard errors are estimated using analytical scores except for the GGEE copula for which the scores are estimated numerically. To obtain the SMLE, we use the Bernstein-Kantorovich sieve with $J_N = 6$ selected by AIC and to obtain the SMLE standard errors we use (10) modeling g using the cosine tensor product sieve with 9 parameters.

Consistency of the FMLE estimator relies on correctness of the assumed copula family. If an incorrect copula family is used in the FMLE, it may be biased (see Prokhorov and Schmidt, 2009). 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 a correct fully parametric model. So we may 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 the above intuition. As expected the FMLE standard errors are mostly smaller than those of QMLE. However, this higher efficiency comes with the potential 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).[14]

4.2 Portfolio Value-at-Risk

Next, we consider an application involving the management of an investment portfolio and show how the use of SMLE can lead to superior estimates of the level of market risk associated with an investment in a portfolio security such as a stock or bond.

Let P_t denote the market price of a security at time t, and let R_t be the associated holding period return between times t and $t+1$. In addition to the rate of expected return $\mu_{R_t} = E[R_t]$ over the holding period, of key interest to an investor in this security are measures that gauge the riskiness of their investment such as variance of the investment return $\sigma^2_{R_t} = E[(R_t - E[R_t])^2]$ and the "Value-at-Risk" (VaR), which given some confidence level α is defined as $V_{1-\alpha}(R_t) = \inf\{R_t : F(R_t;\theta) > 1-\alpha\} = F^{-1}(1-\alpha;\theta)$, for $\alpha \in [0,1]$, where F is the c.d.f. of R_t with parameter vector θ. The 5% VaR, or $V_{0.05}(R_t)$, for example, shows a

[14]Most of the computations for the application were performed using Matlab R2022b on the desktop PC with 6-core Intel(R) i7-8700 CPU @ 3.20GHz and 16GB RAM. QMLE was the fastest and took less than 1 second, FMLEs with analytically defined copulas and densities are also very fast in estimation, less than 1 second, however, deriving analytical expressions for the score required additional time (some were performed in Mathematica). SMLE was more computationally demanding than the analytically-derived FMLEs with 13 seconds to run the estimation and 6 minutes to select the number of sieve elements. The FMLE estimation with the GGEE copula was performed in R 4.2.2 using CopulaOne package by Hua. The procedure took one hour as it heavily relies on numerical methods for integration and differentiation.

level of loss that is only 5% likely to be exceeded between times t and $t+1$, and this measure is widely used in the financial industry to characterize "worst-case" scenarios associated with an investment.

Given a parametric specification for F, a VaR estimate can be easily obtained as $\hat{V}_{1-\alpha}(R_t) = F^{-1}(1-\alpha;\hat{\theta})$, where $\hat{\theta}$ is the estimate of the parameter vector θ. The VaR estimate is therefore a functional of $\hat{\theta}$, and the properties of the estimator for the marginal parameters in $\hat{\theta}$ directly affect the accuracy of the density estimate $\hat{F}(R_t;\hat{\theta})$, and in turn the accuracy of the VaR measure. But this represents an interesting new avenue in VaR analysis since efficiency gains offered by SMLE with respect to the estimation of the marginal parameters in θ may translate into superior estimates of the VaR, without the need to specify the full joint distribution. That is, by using another variable associated with R_t SMLE may lead to a more efficient estimate of θ, and hence a better estimate of the VaR for R_t, while avoiding the risk of biasing $\hat{V}_{1-\alpha}$ due to incorrect choice of a dependence structure which arises in a fully-specified multivariate setting. In the remainder of this section we put this notion to the test.

4.2.1 Estimating 5% VaR for Bank of America Stock

We begin by using SMLE to estimate weekly 5% VaR for a potential investment in the Bank of America (NYSE: BAC) stock. To explore improvements in the BAC VaR estimates arising from SMLE we need to find other variables that are associated with the BAC stock return and therefore contain additional information about BAC returns. To this end, we select BAC trading volume and realized volatility of BAC returns as two such "dependence instruments". The return-volatility and return-volume relationships are both well-documented in the lit-

erature (see, for example, Gervais et al. (2001) and Ang et al. (2006), among others), and we therefore expect that both measures may contain relevant information which we aim to extract using SMLE.

Our sample consists of daily adjusted closing prices for BAC beginning in August of 1999 and ending in December of 2020, and we use Friday closing price P_t to calculate weekly holding period returns as $R_t = \ln(P_t/P_{t-1})$. For each of the weeks we also calculate the change in dollar trading volume during the week as $M_t = \ln(m_t/m_{t-1})$, where $m_t = \sum_j s_j P_j$, with s_j representing the number of shares traded during the j-th day, with the summation being over all trading days of the week. We further estimate the weekly realized volatility V_t of the BAC stock price as the standard deviation of daily returns during the week, and begin by building the marginal models for R_t, M_t and V_t. We select t-Location-Scale distribution to model weekly returns, with the density given by:

$$f_r(R_t; \nu_r, \sigma_r, \mu_r) = \frac{\Gamma\left(\frac{\nu_r+1}{2}\right)}{\Gamma\left(\frac{\nu_r}{2}\right)\sqrt{\pi\nu_r}\sigma_r}\left(1 + \frac{1}{\nu_r}\left(\frac{R_t - \mu_r}{\sigma_r}\right)^2\right)^{-\frac{\nu_r+1}{2}}, \tag{11}$$

where $\Gamma(\cdot)$ is the Gamma function, μ_r and σ_r are the location and scale parameters, and ν_r is the tail parameter which allows for excess kurtosis often present in financial returns. We adopt the same specification, with distinct parameters, for the marginal model of M_t and denote the density of M_t by $f_m(M_t; \nu_m, \sigma_m, \mu_m)$, but chose the Beta distribution to model volatility so that to capture non-negativity and skewness in V_t, with the density given by:

$$g(V_t; \alpha, \beta) = \frac{V_t^{\alpha-1}(1 - V_t)^{\beta-1}}{B(\alpha, \beta)}, \tag{12}$$

where $B = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

Note that in principle the setup of Section 2 is general enough to accommodate any dimension of the data, meaning that we can use SMLE to augment estimation of the marginal parameters μ_r, σ_r, and ν_r using any number of available "dependence instruments". To avoid the curse of dimensionality, here we restrict ourselves to two. To better understand the relative improvements from the addition of our volatility and volume measures with respect to estimation of the BAC weekly VAR, we first add V_t and M_t separately. We build individual bivariate models for (R_t, M_t) and (R_t, V_t) in Section 4.2.2 and study the behaviour of the resulting VaR estimate before considering the case where we add both "instruments" simultaneously and estimate a trivariate model for (R_t, M_t, V_t) in Section 4.2.3. In all three cases, to establish a benchmark we first obtain parameter estimates using QMLE under the assumption of independence between these variables, but then re-estimate the marginal parameters in (11) using FMLE, with a bi-variate t-copula. In addition to correlation, the t-copula captures dependence in the tails of the joint distribution and is a common choice in the financial industry today. Lastly, we obtain a third set of parameter estimates using SMLE. To compare the resulting sets of VaR measures we use likelihood-based scoring rules for comparing density forecasts proposed by Diks et al. (2011).

4.2.2 The case of two dimensions: adding trading volume and return volatility individually

We start by leveraging only the information contained in trading volume and build a bivariate model for (R_t, M_t). First, to establish a benchmark we estimate the marginal parameters in

$F_r(R_t; \nu_r, \sigma_r, \mu_r)$ using QMLE under the assumption of independence between BAC returns and trading volume, which amounts to maximizing log-density given by:

$$\ln h(R_t, M_t; \theta) = \ln f_r(R_t; \nu_r, \sigma_r, \mu_r) + \ln f_m(M_t; \nu_m, \sigma_m, \mu_m) \tag{13}$$

with respect to vector θ containing all parameters in h, or effectively by estimating the marginal parameters (ν_r, σ_r, μ_r) and (ν_m, σ_m, μ_m) using univariate MLE. We then re-estimate the parameter vector using FMLE, where we allow for a degree of correlation as well as tail dependence between BAC returns and changes in the trading volume by selecting the bi-variate t-copula in place of independence. We therefore maximize

$$\begin{aligned} \ln h(R_t, M_t; \theta) &= \ln f_r(R_t; \nu_r, \sigma_r, \mu_r) + \ln f_m(M_t; \nu_m, \sigma_m, \mu_m) \\ &+ c_t(F_r(R_t; \mu_r, \sigma_r, \nu_r), F_m(M_t; \mu_m, \sigma_m, \nu_m); \rho, \tau), \end{aligned} \tag{14}$$

where c_t is the bivariate t-copula log-density parametrized by the correlation coefficient ρ and tail thickness parameter ρ. Lastly, in constructing the copula term in (14) we use the Bernstein-Kantorovich sieve from (2) with $J_N = 5$ and estimate the marginal parameters in (11) one final time, now using SMLE.

To construct the bivariate model for the case where returns are paired with volatility instead of trading volume we repeat these steps but replace the volume density f_m in (14) with g from (12).

We use a historical three-year rolling window to obtain all estimates and calculate corresponding weekly 5% VaR figures for BAC for each of the trading weeks in the sample as

$\hat{V}_{t,0.05}(R_t) = \hat{F}_t^{-1}(0.05; \hat{\nu}_r, \hat{\sigma}_r, \hat{\mu}_r)$, with the marginal parameters $(\hat{\nu}_r, \hat{\sigma}_r, \hat{\mu}_r)$ obtained using QMLE, FMLE, and SMLE applied to the same sample. To compare the behavior of these VaR measures and to gauge economic significance of the differences between QMLE, FMLE, and SMLE-based VaR we focus on two criteria: the number of times actual losses exceed the corresponding VaR estimate (which we refer to as exceedances), and the relative accuracy of the VaR, which we assess using the likelihood-based scoring rule of Diks et al. (2011).

VaR exceedances that occur more frequently than $(1 - \alpha)\%$ of the time may suggest a biased VaR, and avoiding this bias is particularly important for institutional investors such as banks. Many larger banks are subject to capital adequacy requirements that are part of regulatory frameworks, for example the Bank of International Settlements Basel III Accord. Through their compliance process banks must maintain a portion of capital invested in risk-free assets as security against possible trading losses. Such "regulatory capital" acts as a cushion against default in times of extreme market volatility and its size generally depends on the aggregate level of risk associated with the bank's balance sheet. VaR is often used as part of this risk calculation, and banks face financial penalties when the rate of exceedances is too high, and consequently the level of regulatory capital is too low.

We find the rate of exceedances for SMLE VaRs to be virtually the same as that for QMLE and FMLE, suggesting that any differences in the behavior of SMLE VaR are not due to the presence of bias in the VaR measure, and will not come at a negative economic cost to a would-be user relative to QMLE or FMLE.

The accuracy of our VaR forecasts largely depends on the accuracy of underlying density estimates $\hat{F}(R_t; \hat{\nu}_r, \hat{\sigma}_r, \hat{\mu}_r)$ obtained using QMLE, FMLE and SMLE. We follow Diks et al.

(2011) and adopt a censored likelihood-based scoring rule for assessing the accuracy of competing density forecasts in the specified region of interest, which in our case is the left tail of the BAC return distribution. For each of the trading weeks in the sample we calculate censored scores for QMLE, FMLE and SMLE-based density estimates as

$$
\begin{aligned}
S(R_{t+1}; \hat{f}_r) &= w_t(R_{t+1}) \log \hat{f}_r(R_{t+1}; \hat{\mu}_r^t, \hat{\sigma}_r^t, \hat{\nu}_r^t) && (15) \\
&+ (1 - w_t(R_{t+1})) \log \left(1 - \int w_t(s) \hat{f}_r^t(s; \hat{\mu}_r^t, \hat{\sigma}_r^t, \hat{\nu}_r^t) ds \right), && (16)
\end{aligned}
$$

where $\hat{\mu}_r^t, \hat{\sigma}_r^t$, and $\hat{\nu}_r^t$ are the marginal parameters of the return distribution which we estimate using a rolling window ending in period t. The function $w(s)$ weighs observations proportional to their distance from the left tail. In the extreme case $w(s)$ can be defined as an indicator, discarding all returns that fall above a certain threshold. We adopt a specification where $w(s) = 1/(1 + \exp(a(y - s))$, therefore letting all observations along the return spectrum influence the score, while enabling us to assign higher weights to observations belonging to the left tail. For the purpose of weighting we set the return threshold y to negative 8%, which is the fifth percentile of R_t in the whole sample. The parameter a in $w(s)$ determines the rate with which weights diminish with distance from our threshold. We set $a = 30$, and also note that our results do not appear to be sensitive to alternative choices of these parameters, or the weighting function itself.

Figure 4 shows smoothed differences in weekly values of $S(R_{t+1}; \hat{f}_r)$ for SMLE relative to QMLE and FMLE, for the case where returns are paired with volatility. Positive differences indicate higher SMLE scores, and for most of the trading weeks SMLE appears to produce

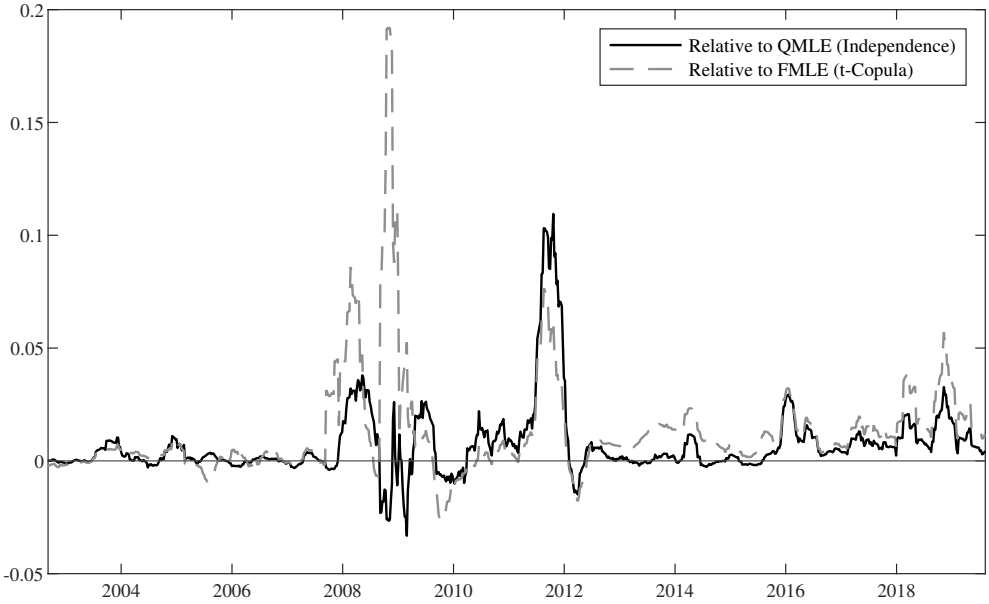


Figure 4: Differences in tail forecast accuracy between SMLE and QMLE, and SMLE and FMLE. Twelve week moving average of score difference, January 2002 - December 2020.

significantly higher scores, meaning more accurate tail forecasts and hence VaR estimates than both QMLE and FMLE. Interestingly, this appears to be the case particularly in times of market turbulence, such as during the sub-prime crisis of 2007-2008 and the European debt crisis of 2011-2012. We also find this to be the case when returns are paired with trading volume M_t instead of volatility in this way.

	Trading volume		Realized volatility	
	SMLE - QMLE	SMLE - FMLE	SMLE - QMLE	SMLE - FMLE
Mean difference	0.0016	0.0029	0.0072	0.0128
t-Ratio	2.2348	4.0483	3.1500	5.6225

Table 4: Differences in mean tail forecast accuracy scores between SMLE and QMLE, SMLE and FMLE.

To formally test for the presence of positive difference between SMLE, QMLE and FMLE tail forecast scores we calculate differences in average scores in the sample and estimate the

standard deviations of the differences using delete-d jack-knife, with $d = 10$. Table 4 shows differences between mean SMLE, QMLE and FMLE scores when returns are paired with realized volatility, and separately with trading volume. As before, a positive score difference indicates greater mean score for SMLE. In all cases, we find that we can reject the null hypothesis of equal mean scores in favour of the alternative of greater SMLE mean score at the 5% significance level.

Superior performance of SMLE-based VaR estimates is potentially significant from a practical risk management standpoint. We speculate that such improvement may be due to better ability of the Sieve copula to capture temporal shifts in market dependence as well as possible asymmetries, but we leave these questions for future work.

4.2.3 The case of three dimensions: combining trading volume with volatility

Our setup allows for simultaneous addition of volatility and trading volume. QMLE in such three-dimensional case amounts to operating on the sum of marginal log-densities, now given by

$$\ln h(R_t, M_t, V_t; \theta) = \ln f_r(R_t; \nu_r, \sigma_r, \mu_r) + \ln f_m(M_t; \nu_m, \sigma_m, \mu_m) + \ln g(V_t; \alpha, \beta). \qquad (17)$$

Selecting a trivariate t copula $c_t(f_r, f_m, g; \Omega, \tau)$, which is now parametrized by a correlation matrix Ω in addition to the tail thickness parameter τ to model dependence between R_t, V_t

and M_t yields log-density for FMLE:

$$\begin{aligned} \ln h(R_t, M_t; \theta) &= \ln f_r(R_t; \nu_r, \sigma_r, \mu_r) + \ln f_m(M_t; \nu_m, \sigma_m, \mu_m) + \ln g(V_t; \alpha, \beta) \\ &+ \ln c_t(F_r(R_t; \mu_r, \sigma_r, \nu_r), F_m(M_t; \mu_m, \sigma_m, \nu_m), G(V_t; \alpha, \beta); \Omega, \tau), \end{aligned} \tag{18}$$

where $G(V_t; \alpha, \beta)$ is the cdf corresponding to the pdf $g(V_t; \alpha, \beta)$. As before, we maximize the log-likelihood based on (18) with respect to parameter vector θ to estimate the model parameters using SMLE, but construct the copula term using the Bernstein-Kantorovich sieve from (2), now with $m = 3$, but keeping $J_N = 5$.

We repeat all steps from the previous section and again follow Diks et al. (2011) in obtaining forecast accuracy scores for SMLE, QMLE and FMLE using the same parameters, but shortening our sample to 2012-2020 period to focus on more recent data and to accommodate increased computational complexity in higher dimensions. Similar to bi-variate setup involving only volatility or trading volume, we find significant improvements in forecast accuracy in the tails arising from the use of SMLE, with the differences in mean scores being positive and statistically significant at the 5% significance level. We summarize these results in Table 5.

	Trading volume and realized volatility	
	Relative to QMLE	Relative to FMLE
SMLE score difference	0.0022	0.0074
t-Ratio	2.3001	5.2096

Table 5: Differences in mean tail forecast accuracy scores between SMLE and QMLE, SMLE and FMLE.

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 substantial. The efficiency bound is determined by the dependence strength and we show that our estimator reaches that bound. We illustrate the usability of SMLE with two empirical applications in insurance and financial risk management. The dependence structure itself is not modeled directly which can be viewed as a drawback in some cases. However, the procedure has clear advantages when the core interest is in estimating features of the marginals whereas dependence is viewed as nuisance parameters.

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 and three-dimensional problems we consider).

In higher dimensions, the application of our approach is limited but a productive way to think about applying it is in the settings where one uses low dimensional copulas to arrive at a high-dimensional likelihood such as vine-copulas, factor copulas or composite densities (see, e.g., Scheffer and Weiss, 2017; Krupskii and Joe, 2013; Anatolyev et al., 2018). For example, Scheffer and Weiss (2017) claim they were able to reach $d = 15$ using vines of bivariate

Bernstein copulas. We leave these approaches for future work.

Simple alternatives to the proposed method include a fully parametric ML estimation problem and various weighting schemes of the QMLE moment conditions (see, e.g., Prokhorov and Schmidt, 2009; Nikoloulopoulos et al., 2011). Although simpler computationally, the weighting schemes usually do not use information beyond correlation of the marginal scores while the full MLE imposes an assumption on the dependence structure, which, if violated, renders the ML estimator inconsistent. Moreover, robust parametric copulas are often robust because they are redundant. So the proposed estimator seems to offer a natural 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 the development of such methods for future work.

6 Supplemental Material

Supplemental materials for this article are available online at

`http://research.economics.unsw.edu.au/vpanchenko/software/scopula.zip`.

They include Matlab codes implementing the Bernstein-Kantorovich sieve and other codes used in simulations and applications, as well as relevant data.

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.

Anatolyev, S., R. Khabibullin, and A. Prokhorov (2018): *Estimating Asymmetric Dynamic Distributions in High Dimensions*, John Wiley & Sons, Ltd, chap. 8, 169–197.

Anderson, E., A. Prokhorov, and Y. Zhu (2021): "A Simple Estimator of Two-Dimensional Copulas, with Applications1," *Oxford Bulletin of Economics and Statistics*, 82, 1375–1412.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006): "The cross-section of volatility and expected returns," *The Journal of Finance*, 61, 259–299.

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

Bouezmarni, T., J. V. Rombouts, and A. Taamouti (2010): "Asymptotic properties of the Bernstein density copula estimator for α-mixing data," *Journal of Multivariate Analysis*, 101, 1–10.

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.

Diers, D., M. Eling, and S. D. Marek (2012): "Dependence modeling in non-life insurance using the Bernstein copula," *Insurance: Mathematics and Economics*, 50, 430–436.

Diks, C., V. Panchenko, and D. Van Dijk (2011): "Likelihood-based scoring rules for comparing density forecasts in tails," *Journal of Econometrics*, 163, 215–230.

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

Gervais, S., R. Kaniel, and D. H. Mingelgrin (2001): "The high-volume return premium," *The Journal of Finance*, 56, 877–919.

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

Godambe, V. and M. Thompson (1978): "Some aspects of the theory of estimating equations," *Journal of Statistical Planning and Inference*, 2, 95–104.

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.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hao, B., A. Prokhorov, and H. Qian (2018): "Moment redundancy test with application to efficiency-improving copulas," *Economics Letters*, 171, 29 – 33.

Hill, J. B. and A. Prokhorov (2016): "GEL estimation for heavy-tailed GARCH models with robust empirical likelihood inference," *Journal of Econometrics*, 190, 18–45.

Hirukawa, M., I. Murtazashvili, and A. Prokhorov (2020): "Uniform Convergence Rates for Non-parametric Estimators Smoothed by the Beta Kernel," *Working Paper*.

Hoff, P. D., X. Niu, and J. A. Wellner (2014): "Information bounds for Gaussian copulas," *Bernoulli*, 20, 604–622.

Horowitz, J. L. (1998): *Semiparametric Methods in Econometrics*, Lecture Notes in Statistics 131, Springer-Verlag New York, 1 ed.

Hu, T., Q. Zhou, and J. Sun (2017): "Regression analysis of bivariate current status data under the proportional hazards model," *Canadian Journal of Statistics*, 45, 410–424.

Hua, L. (2017): "On a bivariate copula with both upper and lower full-range tail dependence," *Insurance: Mathematics and Economics*, 73, 94–104.

Joe, H. (1997): *Multivariate models and multivariate dependence concepts*, CRC press.

——— (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.

Krupskii, P. and H. Joe (2013): "Factor copula models for multivariate data," *Journal of Multivariate Analysis*, 120, 85–101.

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.

Nikoloulopoulos, A. K., H. Joe, and N. R. Chaganty (2011): "Weighted scores method for regression models with dependent data," *Biostatistics*, 12, 653–665.

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

Pitt, M. K. and S. G. Walker (2005): "Constructing Stationary Time Series Models Using Auxiliary Variables with Applications," *Journal of the American Statistical Association*, 100, 554–564.

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.

Scheffer, M. and G. N. F. Weiss (2017): "Smooth nonparametric Bernstein vine copulas," *Quantitative Finance*, 17, 139–156.

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

Segers, J., R. van den Akker, and B. J. M. Werker (2014): "Semiparametric Gaussian copula models: Geometry and efficient rank-based estimation," *The Annals of Statistics*, 42, 1911–1940.

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.

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

White, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica: Journal of the econometric society*, 1–25.

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.

Appendix

Appendix A. Basics of sieve MLE: Compared to other nonparametric methods such as kernels, local linear estimators, etc., the method of linear sieves is quite simple – the infinite dimensional optimization is reduced to a regular parametric MLE.

Given a sequence of approximating spaces Γ_N (sieves), such that $\bigcup_N \Gamma_N$ is dense in Γ, the optimization is done over the finite dimensional 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 to a multivariate copula setting, we can write them in a tensor product form as follows:

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

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. The number of sieve elements in the tensor product 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.

If we write, as in the main text, 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, and let $\theta = (\beta', \gamma)$, then the sieve MLE (SMLE) can be written as follows

$$\hat{\theta} = \arg \max_{\theta \in \Theta_N} \sum_{i=1}^{N} \ln h(\mathbf{y}_i; \theta) \tag{19}$$

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 mentioned above 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 in the main text, a consistent estimator of the SMLE asymptotic covariance matrix can be obtained in some cases using standard MLE.

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 below we provide reference to that approach for readers more familiar with it.)

Assumption A4(C): Denote $l_i(\theta) = \ln h(\mathbf{y}_i; \theta)$, $l(\theta) = \frac{1}{N}\sum_{i=1}^{N} l_i(\theta)$ and $0 < \varepsilon_N = o(N^{-1/2})$. Further let $\mu_N(g) = \frac{1}{N}\sum_{i=1}^{N}[g(\mathbf{y}_i) - Eg(\mathbf{y}_i)]$ for some function $g(\cdot)$ as in Chen et al. (2006).

Assume that, for any $\tilde{\theta} : ||\tilde{\theta} - \theta_o|| = O_p(\delta_N)$ and $v : ||v|| = O(\delta)$, we have

$$E\left[\frac{d\,\dot{l}_i(\tilde{\theta})[\nu]}{d\theta'}[\nu] - \frac{d\,\dot{l}_i(\theta_0)[\nu]}{d\theta'}[\nu]\right] = o(N^{-1}) \tag{20}$$

and

$$\mu_N\left(\frac{d\,l_i(\tilde{\theta})}{d\theta'}[\Pi_N\nu^*] - \frac{d\,l_i(\theta_0)}{d\theta'}[\Pi_N\nu^*]\right) = o_p(N^{-1/2}). \tag{21}$$

For lower level assumptions on individual derivatives, see Assumptions 5 and 6 in Chen et al. (2006). As mentioned there, these assumptions are easily satisfied when marginal densities are twice continuously differ-

entiable around true β and the unknown copula density is in some smooth function class, e.g., Hölder, and is bounded away from zero.

Appendix B. Proofs of Theorems

Proof of Theorem 1: Let $r_i(\theta) = l_i(\theta) - l_i(\theta_0) - \frac{dl_i(\theta)}{d\theta'}[\theta - \theta_0]$. By the definition of $\hat{\theta}$ in (19),

$$\begin{aligned} 0 &\leq l(\hat{\theta}) - l(\hat{\theta} \pm \varepsilon_N \Pi_N \nu^*) = \mu_N(l_i(\hat{\theta}) - l_i(\hat{\theta} \pm \varepsilon_N \Pi_N \nu^*)) + E(l_i(\hat{\theta}) - l_i(\hat{\theta} \pm \varepsilon_N \Pi_N \nu^*)) \\ &= \pm\varepsilon_N \tfrac{1}{N} \textstyle\sum_{i=1}^N \frac{dl_i(\hat{\theta})}{d\theta'} [\Pi_N \nu^*] + \mu_N(r_i(\hat{\theta}) - r_i(\hat{\theta} \pm \varepsilon_N \Pi_N \nu^*)) + E(r_i(\hat{\theta}) - r_i(\hat{\theta} \pm \varepsilon_N \Pi_N \nu^*)) \end{aligned}$$

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

$$\frac{1}{N} \sum_{i=1}^N \frac{dl_i(\theta_o)}{d\theta'} [\Pi_N \nu^* - \nu^*] = o_p(N^{-1/2}) \tag{22}$$

$$E(r_i(\hat{\theta}) - r_i(\hat{\theta} \pm \varepsilon_N \Pi_N \nu^*)) = \pm\varepsilon_N \langle \hat{\theta} - \theta_o, \nu^* \rangle \pm \varepsilon_N \, o_p(N^{-1/2}) \tag{23}$$

$$\mu_N(r_i(\hat{\theta}) - r_i(\hat{\theta} \pm \varepsilon_N \Pi_N \nu^*)) = \varepsilon_N \, o_p(N^{-1/2}) \tag{24}$$

It will then follow that

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

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

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

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 (22)-(24).

Equation (22) holds by Assumption A4(B), since $||\Pi_N \nu^* - \nu^*|| = o(1)$. To show (23), note that, under

Assumption A4(C) Eq.(20),

$$\begin{aligned} Er_i(\theta) &= E\left(l_i(\theta) - l_i(\theta_0) - \frac{dl_i(\theta)}{d\theta'}[\theta-\theta_0]\right) \\ &= \frac{1}{2}E\left[\frac{d\,\dot{l}_i(\tilde{\theta})[\theta-\theta_0]}{d\theta'}[\theta-\theta_0] - \frac{d\,\dot{l}_i(\theta_0)[\theta-\theta_0]}{d\theta'}[\theta-\theta_0]\right] + \frac{1}{2}E\left(\frac{d\,\dot{l}_i(\theta_0)[\theta-\theta_0]}{d\theta'}[\theta-\theta_0]\right) + \varepsilon_N o_p(N^{-1/2}) \\ &= \frac{1}{2}E\left(\frac{d\,\dot{l}_i(\theta_0)[\theta-\theta_0]}{d\theta'}[\theta-\theta_0]\right) + \varepsilon_N o_p(N^{-1/2}) + o_p(N^{-1}), \end{aligned}$$

where $\tilde{\theta}$ is between $\hat{\theta}$ and $\hat{\theta} \pm \varepsilon_N \Pi_N \nu^*$. Therefore, as shown by Chen et al. (2006, proof of Theorem 1),

$$\begin{aligned} E(r_i(\hat{\theta}) - r_i(\hat{\theta} \pm \varepsilon_N \Pi_N \nu^*)) &= -\frac{||\hat{\theta}-\theta_0||^2 - ||\hat{\theta} \pm \varepsilon_N \Pi_N \nu^* - \theta_0||^2}{2} + o_p(N^{-1/2}) + o_p(N^{-1}) \\ &= \pm\varepsilon_N \langle \hat{\theta} - \theta_o, \nu^* \rangle + \varepsilon_N o_p(N^{-1/2}) + o_p(N^{-1}). \end{aligned}$$

To show (24), we note, that under Assumption A4(C) Eq.(21), we have

$$\begin{aligned} \mu_N(r_i(\hat{\theta}) - r_i(\hat{\theta} \pm \varepsilon_N \Pi_N \nu^*)) &= \mu_N\left(l_i(\hat{\theta}) - l_i(\hat{\theta} \pm \varepsilon_N \Pi_N \nu^*) \pm \varepsilon_N \frac{dl_i(\theta_0)}{\theta'}[\Pi_N \nu^*]\right) \\ &= \pm\varepsilon_N \mu_N \left(\frac{dl_i(\tilde{\theta})}{\theta'}[\Pi_N \nu^*] - \frac{dl_i(\theta_0)}{\theta'}[\Pi_N \nu^*]\right) = \varepsilon_N o_p(N^{-1/2}), \end{aligned}$$

where $\tilde{\theta}$ is between $\hat{\theta}$ and $\hat{\theta} \pm \varepsilon_N \Pi_N \nu^*$. This completes the proof.

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 $s_i = \frac{dl_i(\theta_t)}{dt}\Big|_{t=0} = \frac{d \ln h(\mathbf{y}_i;\theta_t)}{dt}\Big|_{t=0} = \frac{d \ln h(\mathbf{y}_i;\theta_o)}{d\theta}[\dot{\theta}]$. 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 $\frac{d\rho(\theta(t))}{dt}\Big|_{t=0} = 1$, 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 (7). Thus, SPLB $= ||\nu^*||^2$.

Appendix C. Implementation algorithm for SMLE using Bernstein-Kantorovich sieve.

We observe an i.i.d. sample $\{\mathbf{y}_i\}_{i=1}^N = \{y_{1i}, \ldots, y_{mi}\}_{i=1}^N$. Assume that the corresponding marginal distributions $F_1(y_1;\beta_1), \ldots, F_m(y_m;\beta_m)$ are known up to their parameters β_js. The dependence is modeled non-parametrically using the Bernstein-Kantorovich sieve (see Section 2.1, which provides definitions and formulas) as follows:

1. Initialization. Use QMLE estimates as initial values, $\hat{\beta}_{\text{init}}$. Compute m marginal CDF transforms of each variable, $u_{ji} = F_j(y_{ji};\hat{\beta}_{j,\text{init}}), j = 1, \ldots, m$. For a given J_N, which determines the number of sieve parameters (we discuss the selection later), use u_{ji}s, to compute empirical m-dimensional histogram of u_{ji}s on $[0,1]^m$ with the grid $J_N \times \cdots \times J_N$, by counting the number of u_{ji}s falling in (3) and dividing

it by the number of observations. Use the computed values as the initial values for $\omega_{\mathbf{v}}$, the parameters of the Bernstein-Kantorovich sieve given in (2).

2. Estimation. The parameters of the marginals, β_js, and the sieve, $\omega_{\mathbf{v}}$s, are jointly estimated by maximizing the log-likelihood $\ln \mathcal{L}(\beta, \omega_{\mathbf{v}}) = \sum_{i=1}^{N} [\sum_{j=1}^{m} \ln f_j(y_{ji}; \beta_i) + \ln c_{J_N}(F_1(y_{1i}) \dots F_m(y_{mi}); \omega_{\mathbf{v}})]$ with the initial values given in steps 1. When estimating $\omega_{\mathbf{v}}$ it is important to impose the uniform marginals restriction, given after (2). In the bi-variate case, the parameters of the sieve can be represented as a $J_N \times J_N$ matrix of non-negative elements, each of whose rows and columns sums to $1/J_N$. In practice, these restrictions are imposed either by using constrained optimization (this works well for relatively small J_N) or by using transformations to ensure that all sieve parameters are positive and sum to 1 and subtracting the quadratic penalty $\mathcal{P} \sum_{\ell} \sum_{v_\ell} (\sum_{\mathbf{v}_{\text{-}\ell} | v_\ell} \omega_{\mathbf{v}} - 1/J_N)^2$ from the log-likelihood. $\mathcal{P}$ determines the tightness of the constraints at the optimum and should be sufficiently high relative to the typical value of the log-likelihood.

3. Selection of J_N. We use AIC $= 2k - \ln \hat{\mathcal{L}}$, where k is the number of free parameters including the sieve parameters and $\hat{\mathcal{L}}$ is the value of likelihood at the maximum. Other selections criteria: likelihood cross-validation gave similar results but was too computationally demanding, BIC selected a higher J_N than optimal (based on simulations).

4. Post-estimation. Asymptotic variance is computed by inverting the Fisher information matrix consistently estimated by a sample average of $S_\beta S_\beta'$ at the value of the parameter estimates, where S_β is defined in (6). The terms g_q^* are estimated by solving (8) and are approximated with the tensor product cosine sieve $g_q^*(\mathbf{u})/c(\mathbf{u}) = \sum_{k_1=1}^{J_N} \cdots \sum_{k_m=1}^{J_N} a_{k_1,\dots,k_m} \cos(k_1 \pi u_1) \times \dots \times \cos(k_m \pi u_m)$. Note that we use the cosine basis with no constant, which ensures that $\int g_q(\mathbf{u})/\hat{c}(\mathbf{u})\, d\mathbf{u} = 0$ holds.