Efficient estimation of parameters in marginals in semiparametric multivariate models*

Valentyn Panchenko† Artem Prokhorov‡

October 19, 2011

Abstract

Recent literature on semiparametric copula models focused on the situation when the marginals are specified nonparametrically and the copula function is given a parametric form. For example, this setup is used in Chen, Fan and Tsyrennikov (2006) [Efficient Estimation of Semiparametric Multivariate Copula Models, JASA] who focus on efficient estimation of copula parameters. We consider a reverse situation when the marginals are specified parametrically and the copula function is modelled nonparametrically. This setting is no less relevant in applications. We use the method of sieve for efficient estimation of parameters in marginals, derive its asymptotic distribution and show that the estimator is semiparametrically efficient. Simulations suggest that the sieve MLE can be up to 70% more efficient relative to QMLE depending on the strength of dependence between the marginals. An application using insurance company loss and expense data demonstrates empirical relevance of this setting.

JEL Classification: C13

Keywords: sieve MLE, copula, semiparametric efficiency

*Helpfull comments of seminar participants at University of Toronto, University of Pittsburgh, University of New South Wales, Concordia University, QMF09, Panel Data 2009 and FESAMES09 are gratefully acknowledged. We thank Professors Frees and Valdez for kindly providing the loss-ALAE data originally collected by the US Insurance Services Office (ISO).

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

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

1 Introduction

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

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

We will use the well known representation of log-joint-density in terms of log-marginal-

densities and the log-copula-density:

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

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

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

It is also well understood that, unlike QMLE, FMLE is generally not robust to copula misspecification. That is, the efficiency gains will come at the expense of an asymptotic bias if the joint density is misspecified. Prokhorov and Schmidt (2009) point out that there

are robust parametric copulas, for which the pseudo MLE (PMLE) based on an incorrectly specified copula family leads to a consistent estimation. But a copula that is robust in one problem may not be robust in another, and some robust copulas are robust because they are redundant. So finding a general class of robust non-redundant parametric copulas is difficult if at all possible.

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

The paper is related to the literature on efficient semiparametric estimation of copula parameters with nonparametric marginals (see, e.g., Chen et al., 2006) and on efficient estimation of nonparametric marginals when the copula is fully known (see, e.g., Segers et al., 2008). More generally, it is related to the literature on sieve-based estimation of models that contain unknown functions (see, e.g., Ai and Chen, 2003; Newey and Powell, 2003). It is

also related to the literature on two-step semiparametric estimation (see, e.g., Newey and McFadden, 1994; Severini and Wong, 1992) and the literature on semiparametric efficiency bounds (see, e.g., Bickel et al., 1993; Severini and Tripathi, 2001; Newey, 1990).

The paper by Chen et al. (2006) considers a problem which is the converse of ours – a sieve MLE estimation when the copula has a known parametric form but the marginals are unknown. In that setting, sieves are employed to approximate univariate marginal densities. We are employing sieves to approximate a multivariate (log-)density. So the main difficulty of our setting is that, in high dimensions, we will suffer from the curse of dimensionality. For low dimensional problems, simulations show efficiency gains of up to 70% over QMLE.

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

2 Sieve MLE

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

finite 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 Chen (2007) for a recent survey of sieve methods.

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

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

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

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

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

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

where $u \in [0,1]$ and $a_k, b_k \in R$. The number of sieve elements in the tensor sieve J_N^m is the smoothing parameter analogous to the bandwidth in a kernel estimation – it sets the quality of the sieve approximation. Note that for the purpose of reducing variance, it is useful in practice to approximate the log-copula density, not the copula density itself.

Since in general there is no analytic solution for the MLE of the sieve coefficients, the

practical implementation of tensor sieves is often complicated. As an alternative we consider using Bernstein polynomials, in particular the Bernstein copula density introduced by Sancetta and Satchell (2004):

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

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

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

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

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

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

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

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

In establishing 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 the SMLE of β 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).

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

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

A common smoothness assumption is to restrict the class of functions to be approximated,

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

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

where $D^\alpha = \frac{\partial^\alpha}{\partial x_1^{\alpha_1} \ldots \partial x_m^{\alpha_m}}, \alpha = \alpha_1 + \ldots + \alpha_m$ is the differential operator, and $|x|_E = (x'x)^{1/2}$ is the Euclidean norm. Then g is said to belong to the Hölder class on $[0,1]^m$, denoted $\Lambda^r([0,1]^m)$. It is also called r-smooth on $[0,1]^m$. Most of commonly used densities, including copulas, belong to this class, and various linear sieves, as well as the Bernstein polynomial sieve, are known to approximate such functions well.

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

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

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

Similarly, define

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

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

Since $\rho(\theta) = \lambda'\beta$ is linear on $\bar{V}$, 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 $\rho(\theta)$ in our setting, i.e. $\sup_{0\neq\nu\in\bar{V}} \frac{|\dot{\rho}(\theta_o)[\nu]|}{||\nu||} < \infty$. This will be the case if and only if $\sup_{\nu\neq 0,\nu\in\bar{V}} \frac{|\lambda'\nu_\beta|^2}{||\nu||^2} < \infty$. So we now show when this condition holds.

Similar to Ai and Chen (2003) and Chen et al. (2006), for each component β_q, $q = 1,\ldots,p$, we denote by g_q^* the solution to

$$\inf_{g_q} E\left[\sum_{j=1}^{m}\left\{\frac{\partial \ln f_j(y_j,\beta_o)}{\partial\beta_q} + \left(\frac{1}{c(\mathbf{u})}\frac{\partial c(\mathbf{u})}{\partial u_j}\right)\Big|_{u_k=F_k(y_k,\beta_o)} \frac{\partial F_j(y_j,\beta_o)}{\partial\beta_q}\right\} + \left(\frac{1}{c(\mathbf{u})}g_q(u_1,\ldots,u_m)\right)\Big|_{u_k=F_k(y_k,\beta_o)}\right]^2. \tag{9}$$

Now let $g^* = (g_1^*, \ldots, g_p^*)$. We find the sup by writing

$$\begin{aligned} \sup_{\nu \neq 0, \nu \in \bar{V}} \frac{|\lambda' \nu_\beta|^2}{||\nu||^2} &= \sup_{\nu \neq 0, \nu \in \bar{V}} \left\{ |\lambda' \nu_\beta|^2 \left(E\left[\dot{l}(\theta_o)[\nu]^2 \right] \right)^{-1} \right\} \\ &= \lambda' \left(E S_\beta S_\beta' \right)^{-1} \lambda, \end{aligned} \tag{10}$$

where

$$\begin{aligned} S_\beta' &= \sum_{j=1}^m \left\{ \frac{\partial \ln f_j(y_j, \beta_o)}{\partial \beta'} + \left(\frac{1}{c(\mathbf{u})} \frac{\partial c(u_1, \ldots, u_m)}{\partial u_j} \right) \Big|_{u_k = F_k(y_k, \beta_o)} \frac{\partial F_j(y_j, \beta_o)}{\partial \beta'} \right\} \\ &\quad + \left(\frac{1}{c(\mathbf{u})} g^*(u_1, \ldots, u_m) \right) \Big|_{u_k = F_k(y_k, \beta_o)} \end{aligned} \tag{11}$$

So $\rho(\theta) = \lambda' \beta$ in bounded if and only if $E S_\beta S_\beta'$ is finite and positive definite. This condition can also be interpreted as a local identification condition on β_o.

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

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

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

and the norm of the functional defined as

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

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

Application of the theorem to $\dot{\rho}(\theta_o)[\nu] = \rho(\nu)$ suggests that there exists the Riesz representer $\nu^* \in \bar{V}$ of this functional such that $\lambda'(\hat{\beta} - \beta_o) = \langle \hat{\theta} - \theta_o, \nu^* \rangle$ 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 – a fact which will be useful in the proof of 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.

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

$$\nu^* = \left(I, g^{*'} \right)' \left(E S_\beta S_\beta' \right)^{-1} \lambda \tag{12}$$

It is easy to verify that

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

so the required condition holds.

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 estimation 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 assumptions made in Assumption (2) but we state it separately to simplify the proof.

Assumption 4 *(convergence of sieve MLE and smoothness of higher order term in Taylor expansion) Assume (A) that* $||\hat{\theta}-\theta_o|| = O_P(\delta_N)$ *for* $(\delta_N)^w = o(N^{-1/2})$, $w > 1$ *and there exists* $\Pi_N\nu^* \in V_N - \{\theta_o\}$ *such that* $\delta_N||\Pi_N\nu^* - \nu^*|| = o(N^{-1/2})$ *and (B) that, for any* $\theta : ||\theta - \theta_o|| = O_p(\delta_N)$, *the expected directional derivative* $E\frac{d\dot{l}(\theta)[\nu]}{d\theta'}[\nu] \leq ||\nu||^2$.

A discussion of convergence rates of different sieves is provided by Chen (2007) and references therein; general results on convergence rates of sieve MLE can be found in Wong and Severini (1991); Shen and Wong (1994). Basically, Assumption 4 covers all commonly encountered sieve convergence rates. For example, for the trigonometric sieve, Shen (1997) shows that $||\hat{\theta} - \theta_o|| = O_p(N^{-r/(2r+1)})$, where r is the Hölderian exponent from (8); Ghosal (2001) provides results on convergence rates of the Bernstein sieve.

We can now state our main results.

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

Theorem 2 *Under Assumptions 1-4,* $||\nu^*||^2$ *is the lower bound for semiparametric estimation of* $\lambda'\beta$, *i.e.* $\hat{\beta}$ *is semiparametrically efficient.*

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

minimization problem as follows

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

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

A simpler alternative estimator of the sieve MLE asymptotic variance is provided by Ackerberg et al. (2009). They show that one can use the upper left $p \times p$ block of the usual MLE covariance matrix as an estimate of $(E[S_\beta S_\beta'])^{-1}$ provided that the outer-product-of-the-score form of the covariance matrix is used.

3 Simulations

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

One of the practical problems we face is the choice of the degree of polynomials J_N in finite samples. While some asymptotic results on the rate of convergence and its dependence on J_N are available, they are not informative in the finite sample situation. The literature on sieves suggest using typical model selection techniques, such as BIC, AIC or data driven methods

such as cross-validation. However, the theoretical implications of using these techniques in the context of sieves are not explored.

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

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

Table 1 contains simulation results. MSE is minimized at $J_N = 16$. Thus we are estimating 256 nuisance parameters in the sieve and 2 parameters of the marginals. To overcome the problem of large number of parameters we also considered the case where we restrict some of the parameters of the Bernstein polynomial. The parameters of the Bernstein polynomial are directly related to histogram density estimator. In case when the latter is equal to 0 we restrict the parameters of the Bernstein polynomial to zero. MSE is minimized at $J_N = 17$,

however the effective number of parameters which needs to be estimated is about 100 (it varies for each simulation run).

The optimization is complicated by the restrictions on sieve parameters and parameters of the marginals. We used standard constrained maximization routine in Matlab. We used 1,000 simulation runs. We report the simulated mean of the Sieve MLE, QMLE and Full MLE estimators, their simulated variance and mean square error (MSE), as well as the simulated relative efficiency (RE) and relative MSE (RMSE) of the QMLE with respect to the Sieve MLE. RE is the ratio of the SMLE simulated variance to that of QMLE. RMSE is the ratio of the SMLE simulated MSE to that of QMLE.

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

	μ_1 SMLE	QMLE	FMLE	μ_2 SMLE	QMLE	FMLE
Unrestricted sieve $J_N = 16$						
Mean	0.494139	0.500114	0.500060	0.493562	0.499851	0.499928
Var	0.000087	0.000264	0.000017	0.000088	0.000257	0.000017
MSE	0.000121	0.000264	0.000017	0.000129	0.000257	0.000017
RE	0.329545		0.064394	0.342412		0.066148
RMSE	0.459641		0.064404	0.503645		0.066162
Restricted sieve $J_N = 17$						
Mean	0.494509	0.500114	0.500060	0.494130	0.499851	0.499928
Var	0.000079	0.000264	0.000017	0.000082	0.000257	0.000017
MSE	0.000109	0.000264	0.000017	0.000116	0.000257	0.000017
RE	0.299242		0.064394	0.319066		0.066148
RMSE	0.413431		0.064404	0.453101		0.066162

The result suggests that in this specific situation we were able to improve the efficiency relatively to the QMLE substantially. The efficiency gain was as high as 66-70%. It appears that there is some evidence of downward bias in the estimates based on Sieve MLE for $J_N = 16$.

In the restriced sieve case with $J_N = 17$ the bias seems to become smaller and the variance is also improved. Note that this case corresponds to extremely high negative dependence between the marginals. In simulations using a weaker dependence, the improvements were not as substantial.

4 Application from insurance

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

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

$$F_j(Y_j) = 1 - \left(\frac{\lambda_j + Y_j}{\lambda_j} \right)^{-\theta_j}, \quad j = 1, 2. \tag{13}$$

Interest lies in efficient estimation of the marginal distribution parameters $(\lambda_1, \theta_1, \lambda_2, \theta_2)$, making efficient use of the dependence between the claim amount and ALAE. The estimates can be used in pricing insurance, for example.

Additional complications arise due to censoring of Y_1. The likelihood contributions will be different depending on whether the observation is censored or not. Denote the marginal pdfs by $f_j(y_j), j = 1, 2$. The QMLE log-likelihood contribution of an uncensored observation is $\ln f_j(y_j)$. For a censored observation, the contribution is $\ln(1 - F_1(y_1)) = \theta_1(\ln(\lambda_1) - \ln(\lambda_1 + y_1))$. Thus, the QMLE 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. Let $H(y_1, y_2)$ and $h(y_1, y_2)$ denote the joint cdf and pdf, respectively. The FMLE log-likelihood contribution of an uncensored observation is $\ln h(y_1, y_2) = \ln f_1(y_1) + \ln f_2(y_2) + \ln c(F_1(y_1), F_2(y_2))$. To derive the contribution of a censored observation we follow Frees and Valdez (1998) in observing that $Prob(Y_1 \geq y_1, Y_2 \leq y_2) = F_2(y_2) - H(y_1, y_2)$. So the log-likelihood contribution of a censored observation is $f_2(y_2) - H_2(y_1, y_2)$, where $H_2(y_1, y_2) = \frac{\partial H(y_1, y_2)}{\partial y_2}$. But $H(y_1, y_2) = C(F_1(y_1), F_2(y_2))$ so $H_2(y_1, y_2) = C_2(F_1(y_1), F_2(y_2)) f_2(y_2)$, where $C_2(u_1, u_2) = \frac{\partial C(u_1, u_2)}{\partial u_2}$. Therefore the full log-likelihood contribution for observation i can be written as

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

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

numerically.

The extra term will carry over to the variance problem (9) and a consistent estimate of the SMLE variance, $\hat{V}$, will now be

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

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

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

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

Table 2: QMLE, SMLE, Frank copula based FMLE for insurance application with standard errors

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

5 Concluding Remarks

We have proposed an efficient semiparametric estimator of marginal distribution parameters. This is a sieve maximum likelihood estimator based on a finite-dimensional approximation

of the unspecified part of the joint distribution. As such, the estimator inherits the costs and benefits of the multivariate sieve MLE. The major benefit permitted by sieve MLE is the increased relative asymptotic efficiency compared to quasi-MLE. We showed that the efficiency gains are non-trivial. In some simulations the relative efficiency with respect to QMLE was as low as 0.3 – a 70% improvement.

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

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

References

Ackerberg, D., X. Chen, and J. Hahn (2009): "A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators," *UCLA Working Paper*.

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

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1993): *Efficient and adaptive estimation for semiparametric models*, Baltimore: Johns Hopkins University Press.

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

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

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

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

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

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

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

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

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

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

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

Newey, W. and D. McFadden (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, vol. IV, 2113–2241.

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

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

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

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

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

Severini, T. A. and G. Tripathi (2001): "A simplified approach to computing efficiency bounds in semiparametric models," *Journal of Econometrics*, 102, 23–66.

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

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

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

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.

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

6 Appendix

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

Under Assumption 2, $l(\theta)$ is twice continuously differentiable w.r.t. t and

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

By the definition of $\hat{\theta}$ in (7) and the Taylor expansion,

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

Below we show that

$$\frac{1}{N}\sum_{i=1}^N \frac{dl_i(\theta_o)}{d\theta'}\,[\Pi_N \nu^* - \nu^*] = o_p(N^{-1/2}) \tag{14}$$

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

$$E\,\frac{d^2\, l_i(\theta(s))}{d\theta' d\theta}\,[\pm\varepsilon_N \Pi_N \nu^*, \pm\varepsilon_N \Pi_N \nu^*] = \pm\varepsilon_N \langle \hat{\theta} - \theta_o, \nu^* \rangle \pm \varepsilon_N\, o_p(N^{-1/2}) \tag{15}$$

and

$$\frac{1}{N}\sum_{i=1}^{N}\frac{d^2\,l_i(\theta(s))}{d\theta' d\theta}\,[\pm\varepsilon_N\Pi_N\nu^*,\pm\varepsilon_N\Pi_N\nu^*]-E\,\frac{d^2\,l_i(\theta(s))}{d\theta' d\theta}\,[\pm\varepsilon_N\Pi_N\nu^*,\pm\varepsilon_N\Pi_N\nu^*]=\varepsilon_N\,o_p(N^{-1/2})\;\;(16)$$

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

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

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 (14)-(16).

Equation (14) holds by Assumption 4(A), since $||\Pi_N\nu^*-\nu^*||=o(1)$. To show (15), note that, under Assumption 4(B), uniformly over $\theta(s)$ in a neighborhood of θ_o with $||\theta(s)-\theta_o||=O(\delta_N)$,

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

But, by Assumption (4)(A), $\langle\hat{\theta}-\theta_o,\Pi_N\nu^*-\nu^*\rangle=o_p(N^{-1/2})$. Thus, $\pm\varepsilon_N\langle\hat{\theta}-\theta_o,\Pi_N\nu^*\rangle=\pm\varepsilon_N\langle\hat{\theta}-\theta_o,\nu^*\rangle\pm\varepsilon_N o_p(N^{-1/2})$. For showing (16), recall that

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

Now, for some $\theta(\tau)$ between $\hat{\theta}$ and $\hat{\theta}\pm\varepsilon\Pi_N\nu^*$, write $l_i(\hat{\theta}\pm\varepsilon_N\Pi_N\nu^*)-l_i(\hat{\theta})=\pm\varepsilon_N\frac{dl_i(\theta(\tau))}{d\theta'}[\Pi_N\nu^*]$. Then, the

left hand side of (16) can be written as

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

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

Proof of Theorem 2: We apply the method of Severini and Tripathi (2001). To make it easier to follow for those who know their method, we use their notation and also specify our equivalents of their objects. For some $t_o > 0$ let $\theta(t)$ denote a curve from $[0, t_o]$ into Θ such that $\theta(0) = \theta_o$. The curve we consider is $\theta(t) = \theta_o + t\nu$, for any $\nu \in V$. Let $\dot{\theta}$ denote the slope of $\theta(t)$ at $t = 0$, i.e. $\dot{\theta}$ is tangent to the set Θ at θ_o. For our case, $\dot{\theta} = \nu$. Let $T(\Theta, \theta_o)$ denote the collection of all such tangents $\dot{\theta}'s$ and let $\bar{T}(\Theta, \theta_o)$ denote the linear closure of $T(\Theta, \theta_o)$, i.e. 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 credited for being first to observe that the efficiency bound is 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 observation 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 $\hat{t}$, $AV(\sqrt{N}\rho(\theta(\hat{t})) - \rho(\theta_o)) = AV(\sqrt{N}\hat{t}) \geq ||\nu||^{-2}$. 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 (12). Thus, SPLB $= ||\nu^*||^2$.