Sparse Sieve MLE

Di Liu* **Artem Prokhorov†**

March 2014

Abstract

The Dantzig selector is traditionally used for point estimation by least squares when the number of parameters exceeds the number of observations. This paper uses it to obtain smaller standard errors in a sieve maximum likelihood estimation in a panel setting. We assume correctly specified likelihood-based models for each cross section and the Bernstein polynomial serves as a copula sieve capturing dependence between them. This estimator has smaller standard errors asymptotically than the conventional QMLE but, in finite samples, the number of parameters in the sieve is close to the sample size and may exceed it. At the same time, most of the sieve parameters are close to zero. We propose an estimator that uses the Dantzig selector to find the sparsest vector of the sieve parameters satisfying the first order conditions of the MLE up to a given tolerance level. We show in simulations that our estimator produces a sparse sieve MLE with finite-sample properties very similar to the non-sparse alternative, and substantially better the QMLE. Thus the sparsity imposed by the Dantzig selector is innocuous with respect to the non-asymptotic behavior of the sieve MLE; it also permits a substantial increase in computational efficiency compared to the unrestricted sieve MLE. As a theoretical motivation for the good performance of sparse SMLE, we provide an oracle inequality relating the risk of the sparse estimator with that of an infeasible estimation where an oracle tells us which coefficients are insignificant. We also study the parameter path behavior for various tolerance levels and consider a version of a double Dantzig selector which resolves the arbitrariness in choosing the tolerance level.

JEL Codes: C13
Key Words: Dantzig Selector, Sieve MLE, copula, panel data

*Department of Economics, Concordia University, Montreal; email: flyingliudi@gmail.com
†University of Sydney Business School, Sydney; email: artem.prokhorov@sydney.edu.au

1 Introduction

The Dantzig selector (DS) was recently introduced to deal with linear regressions in which the number of parameters is very large, possibly larger than the number of observations, but some parameters are believed to be zero – a setting known as a sparsity scenario (Candes and Tao, 2007). DS is attractive because of its property – known as the oracle inequality – to achieve a loss very similar to what we would get if we were told (by an oracle) which elements of the true parameter vector are zero (see, e.g., Koltchinskii, 2009). Unlike the LASSO estimator, which shares similar oracle properties, DS gives parameter estimates with the smallest l_1 norm and is computationally simpler because it reduces to a linear programming problem (see, e.g., Bickel, Ritov, and Tsybakov, 2009).

In this paper we consider using DS in a semiparametric sieve maximum likelihood estimation (SMLE) under a sparsity scenario. Basically, we employ DS in an adaptive nonparametric copula density estimation where the number of sieve parameters is potentially larger than the sample size but the sieve parameter space is sparse. Therefore, this work is related to the sparse density estimation via l_1 penalization (SPADES) of Bunea, Tsybakov, Wegkamp, and Barbu (2010), who consider a LASSO-type penalized objective function. Instead, we use the DS approach, minimizing the l_1 norm of the parameter vector directly.

The goal is to use the nonasymptotic nature of the oracle inequalities to achieve in finite samples what SMLE achieves only asymptotically – an estimator that dominates the conventional, independence-based QMLE. In other words, the primary purpose of using DS here is relative efficiency and improved finite sample properties, not model selection.

2 Copula-Based SMLE of Parameters in Marginals

2.1 SMLE and QMLE

Consider the setting of a panel with T time periods and N individuals. Assume T is fixed and $N \to \infty$. We will fix $T = 2$ for simplicity. Suppose that for each cross section, we have a correctly specified parametric likelihood-based model and we can estimate this model consistently using only the cross sectional data. However, it is usually possible to use the entire panel to obtain more efficient estimators (see, e.g., Prokhorov and Schmidt, 2009; Amsler, Prokhorov, and Schmidt, 2013).

The estimator we consider is the sieve MLE (SMLE) (see Chen, 2007, for a review). In essence, this is a maximum likelihood estimator which uses a sieve approximator to the true joint log density. Specifically we follow Panchenko and Prokhorov (2013) and consider a sieve approximator of the copula corresponding to the joint density. In this setting, the SMLE attempts to use information contained in the dependence structure between cross sections.

Let $f(y_{it};\beta), t = 1, 2$, denote the marginal densities for each cross section, indexed by parameter β. Let $h(y_{i1}, y_{i2};\beta)$ denote the joint density of $(y_{i1,}, y_{i2})$ and let $c(u_1, u_2)$ denote the copula density, corresponding to $h(y_{i1,}, y_{i2};\beta)$. We are interested in estimation of β – a parameter vector that collects all unknown parameters from the likelihood-based models for the cross sections. By a well known result due to Sklar (1959),

$$\ln h(y_{i1,}, y_{i2};\beta) = \ln f(y_{i1};\beta) + \ln f(y_{i2};\beta) + \ln c(F(y_{i1};\beta), F(y_{i2};\beta)), \tag{1}$$

where $F(y_{it};\beta)$ denotes the corresponding marginal cdf's. They may be distinct but we will put this aside for the moment.

The SMLE replaces the last term in (1) with a truncated infinite series representation (a sieve) of the copula log density and then carries out the usual optimization over both β and the parameters of that representation. This produces the sieve MLE estamator $\hat{\beta}$. Panchenko and Prokhorov (2013) derive the semiparametric efficiency bound for estimation of β and show that $\hat{\beta}$ achieves it.

Denote the vector of sieve parameters by γ and the sieve approximator by $\ln c_\gamma$. Then, the SMLE maximizes the approximate joint log likelihood

$$\ln L_\gamma(\beta) = \sum_{i=1}^{N} \left[\ln f(y_{i1};\beta) + \ln f(y_{i2};\beta) + \ln c_\gamma(F(y_{i1};\beta), F(y_{i2};\beta))\right] \tag{2}$$

The fundamental logic of the sieve estimation is that when the space of functions to be approximated is not too complex and the approximation error goes to zero sufficiently fast we obtain a $\sqrt{N}$-consistent estimator of β (see, e.g., Shen and Wong, 1994; Shen, 1997).

As an alternative we consider the conventional QML estimator which maximizes the quasi-log-likelihood

$$\ln L^Q(\beta) = \sum_{i=1}^{N} \left[\ln f(y_{i1}) + \ln f(y_{i2};\beta)\right]$$

– identicial to the joint log-likelihood under the assumption of independence between y_{i1} and y_{i2}. It is now well understood that the QMLE is consistent for β but the robust, or "sandwich", version

of the variance matrix should be used if there is dependence between the cross sections.

The last term in $\ln L_\gamma(\beta)$ is what distinguishes SMLE from QMLE. We have assumed that the marginals are correctly specified so the marginal score function – the derivative of $\ln f(y_{it}, \beta)$ with respect to β – is zero mean for both cross sections. Correspondingly, the estimator that maximizes $\ln L_\gamma(\beta)$ requires that the copula score is mean zero while the QMLE requires that it is exactly zero, or equvalently, that the copula is the indepedence copula $c(u, v) = 1$. That is, unlike QMLE, the SMLE implies that the following first-order condition holds:

$$\sum_{i=1}^{N} \nabla_{(\beta,\gamma)} \ln c_\gamma \left(F(y_{i1}; \beta), F(y_{i2}; \beta) \right) = 0$$

We will use this condition in constructing our new estimator.

2.2 Bernstein Polynomial Sieve

Let $[0, 1]^2$ denote the unit cube in $\mathbb{R}^2$. For a distribution function $P_c : [0, 1]^2 \to \mathbb{R}$, a bivariate Bernstein polynomial of order $\mathbf{k} = (k_1, k_2)$ associated with P_c is defined as

$$B_{\mathbf{k},P_c}(\mathbf{u}) = \sum_{j_1=0}^{k_1} \sum_{j_2=0}^{k_2} P_c \left(\frac{j_1}{k_1}, \frac{j_2}{k_2} \right) q_{j_1 k_1}(u_1) q_{j_2 k_2}(u_2) \tag{3}$$

where $\mathbf{u} = (u_1, u_2) \in [0, 1]^2$, $q_{j_s k_s}(u_s) = \binom{k_s}{j_s} u_s^{j_s} (1 - u_s)^{k_s - j_s}$. The polynomial is dense in the space of distribution functions on $[0, 1]^2$ and its order $\mathbf{k}$ controls the smoothness of $B_{\mathbf{k},P_c}$, with a smaller k_s associated with a smoother function along dimension s. Moreover, with the conditions $P_c(0, 1) = P_c(1, 0) = 0$ and $P_c(1, 1) = 1$, $B_{\mathbf{k},P_c}(\mathbf{u})$ is a copula function and is referred to as the Bernstein copula associated with P_c. As $\min\{\mathbf{k}\} \to \infty$, $B_{\mathbf{k},P_c}(\mathbf{u})$ converges to P_c at each continuity point of P_c and if P_c is continuous then the convergence is uniform on the unit cube $[0, 1]^2$ (Sancetta and Satchell, 2004; Zheng, 2011).

The derivative of (3) is the bivariate Bernstein density function

$$\begin{aligned} b_{\mathbf{k},P_c}(\mathbf{u}) &= \frac{\partial^2}{\partial u_1 \partial u_2} B_{\mathbf{k},P_c}(\mathbf{u}) \\ &= \sum_{j_1=1}^{k_1} \sum_{j_2=1}^{k_2} w_{\mathbf{k}}(\mathbf{j}) \prod_{s=1}^{2} \beta(u_s; j_s, k_s - j_s + 1) \end{aligned} \tag{4}$$

where, for $\mathbf{j} = (j_1, j_2)$, $w_{\mathbf{k}}(\mathbf{j}) = \mathbf{\Delta} P_c \left(\frac{j_1 - 1}{k_1}, \frac{j_2 - 1}{k_2} \right)$ are weights derived using the forward differ-

ence operator $\mathbf{\Delta}$, and $\beta(\cdot;\gamma,\delta)$ denotes the probability density function of the β-distribution with parameters γ and δ.

In order to give a mixing interpretation to $w_{\mathbf{k}}$, let Cube$(\mathbf{j},\mathbf{k})$ denote a cube given by $((j_1-1)/k_1, j_1/k_1] \times ((j_2-1)/k_2, j_2/k_2]$ with the convention that if $j_s = 0$ then the interval $((j_s-1)/k_s, j_s/k_s]$ is replaced by the point $\{0\}$. Then, the mixing weights $w_{\mathbf{k}}(\mathbf{j})$ are the probabilities of Cube$(\mathbf{j},\mathbf{k})$ under P_c. The Bernstein density function $b_{\mathbf{k},P_c}(\mathbf{u})$ can thus be viewed as a mixture of beta densities, and if P_c is a copula, $b_{\mathbf{k},P_c}(\mathbf{u})$ is itself a copula density.

Alternatively, if we interpret P_c as an empirical copula on $\left[\frac{1}{k_1}, \frac{2}{k_1}, ..., \frac{k_1}{k_1}\right] \times \left[\frac{1}{k_2}, \frac{2}{k_2}, ..., \frac{k_2}{k_2}\right]$ then $b_{\mathbf{k},P_c}(\mathbf{u})$ can be viewed as a smoothed copula histogram using β-densities as smoothing functions.

The Bernstein copula density has several attractive properties as a sieve for the space of copula densities, which makes it preferrable to other types of sieve. Being a mixture of (a produce of) β-densities, it assigns no weights outside $[0,1]^2$ and it easily extends to dimensions higher than two. Other sieves known to approximate well smooth functions and densities on $\mathbb{R}$ are often subject to the boundary problem and do not extend easily to multivariate settings (see, e.g., Chen, 2007; Bouezmarni and Rombouts, 2010). The Bernstein sieve is a copula density by construction; at the same time, it does not impose symmetry, contrary to other conventional kernels used in mixture models such as multivariate Gaussian (see, e.g., Burda and Prokhorov, 2013).

Most importantly, as a density corresponding to $B_{\mathbf{k},P_c}(\mathbf{u})$, $b_{\mathbf{k},P_c}(\mathbf{u})$ converges, as $\min\{\mathbf{k}\} \to \infty$, to $p_c(\mathbf{u}) \equiv \frac{\partial^2}{\partial u_1 \partial u_2} P_c(\mathbf{u})$ at every point on $[0,1]^2$ where $p_c(\mathbf{u})$ exists, and if p_c is continuous and bounded then the convergence is uniform (Lorentz, 1986). Uniform approximation results for the univariate and bivariate Bernstein density estimator can be found in Vitale (1975) and Tenbusch (1994).

In what follows we will assume $P_c(\mathbf{u})$ to be a continuous copula. As a result, we will omit subscript P_c and let $b_{\mathbf{k}}(\mathbf{u})$ simply denote the Bernstein copula density with weights w_j, where $j = 1, \ldots, J$, indexes the set $\{j_1, j_2\}$. Consequently, we can write the copula density as follows

$$b_{\mathbf{k}}(\mathbf{u}) = \sum_{j}^{J} w_j g_j(\mathbf{u}),$$

where $g_j(\mathbf{u}) = \prod_{s=1}^{2} \beta(u_s; j_s, k_s - j_s + 1)$.

2.3 SMLE with Dantzig Selector

In practice, the SMLE involves a truncation of the Bernstein polynomial approximation at some large values $\mathbf{k}_N \equiv (k_1^*, k_2^*)$. This means there is a large but finite number – possibly different in each coordinate – of the mixing weights w_j in the Bernstein copula density. Let γ_N contain all such mixing weights. Then, $J = \dim\{\gamma_N\} = k_1^* k_2^*$ and it will grow exponentially as we add dimensions. An important issue in adaptive estimation of such models is how to reduce the dimension of γ_N.

2.3.1 Dantzig Selector

The Dantzig selector is an "automatic" mechanism for selecting non-zero parameters in highly parameterized problems. It is "automatic" because we do not need to even set the maximum number of non-zero parameters. So long as there are zero and non-zero elements in the parameter vector, that is, so long as a sparsity scenario applies, the method will pick the non-zero parameters correctly.

The initial application of the Dantzig selector was in linear regressions with more regressors than observations. Suppose we have the following regression model $y = X\theta + u$, where $\theta \in \mathbb{R}^p$, $u \sim N(0, \sigma^2\mathbb{I})$ and X is a $N \times p$ data matrix with possibly fewer rows than columns, i.e. with $N < p$. Then, the Dantzig selector of Candes and Tao (2007) is the solution to the following problem

$$\min_{\theta} ||\theta||_{l_1} \text{ subject to } ||X'(y - X\theta)||_{l_\infty} \leq \lambda_p \, \sigma, \tag{5}$$

where $||\theta||_{l_1} = \sum_{j=1}^{p} |\theta_j|$ is the l_1-norm of θ, $||Z||_{l_\infty} = \max\{|Z_1|, \ldots, |Z_p|\}$ is the l_∞-norm of any vector $Z \in \mathbb{R}^p$, and λ_p is a positive number – a function of p only. Compared to the usual OLS, the Dantzig selector searches for a θ which has the smallest l_1-norm and, within a fixed tolerance level λ, satisfies the normal equations. Beacuse it produces sparse coefficient estimates, it can be used for model selection. For $\lambda = 0$, it reduces to standard OLS.

It is well known (see, e.g., Bickel, Ritov, and Tsybakov, 2009) that this problem can be viewed as a penalized LS problem, written as follows

$$\min_{\theta} \left\{ \text{SSE}(\theta) + 2\lambda_p \sigma \sum_{j=1}^{p} |\theta_j| \right\}, \tag{6}$$

where $\text{SSE}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - X_i\theta)^2$ and the penalty term grows with complexity of θ as measured by the l_1-norm. So the Dantzig selector solves this problem for a vector having the smallest l_1-norm.

The most attractive theoretical property of the Dantzig selector is that there is a nonasymptotic bound on the error in the estimator of θ that is within a factor of $\log p$ of the error achieved if the true predictors are assumed known. To see this, let $\hat{\theta}$ denote the solution. Candes and Tao (2007) show that under certain conditions on X and under a sparsity scenario (which roughly amounts to an identification condition in this model), the following holds with a large probability,

$$||\hat{\theta} - \theta||_{l_2}^2 \leq \text{const} \cdot \lambda_p^2 \cdot \left(\sigma^2 + \sum_{j=1}^{p} \min \left\{ \theta_j^2, \sigma^2 \right\} \right), \tag{7}$$

where $||\theta||_{l_2} = \sqrt{\theta'\theta}$ and λ_p^2 is of order $O(\log p)$.

Now consider a standard LS estimator in the situation when we know (from an oracle) which θ_j's are significant (i.e., larger than the noise, $|\theta_j| > \sigma$). In this case, we can set equal to zero all the elements of θ that are smaller than σ in magnitude and let the OLS estimate the significant elements. If, for simplicity, X is assumed to be the identity matrix, then the MSE of the LS estimate of θ will contain terms equal to σ^2 for each significant θ_j and terms equal to θ_j^2's for each insignificant θ_j's (i.e., for the coordinates within the noise level). That is, the MSE of this infeasible estimator can be written as follows

$$\mathbb{MSE}_{\text{OLS}} = \sum_{i=1}^{p} \min\{\theta_j, \sigma^2\}$$

When we relax the assumption that X is identity but still allow the oracle to tell us which subset of θ_j's is right to use in the OLS, the MSE will be different. However, Candes and Tao (2007) show that, under certain assumptions on X, $\mathbb{MSE}_{\text{OLS}}$ can still be viewed as a proxy for the MSE in the more general setting, which has the following natural interpretation

$$\sum_{i=1}^{p} \min\{\theta_j, \sigma^2\} = \min_{S \subset \{1,\ldots,p\}} ||\theta - \theta_S||_{l_2}^2 + |S|\sigma^2,$$

where S indexes the set of significant θ_j's, θ_S contains θ_j's if j is in S and 0's otherwise and $|S|$ denotes the number of non-zero elements in S. Of course, the first term of this representation is the squared bias of the ideal estimator and the second is its variance

So the DS nearly achieves the MSE of the ideal estimation, in which an oracle tells us the

composition of S. Specifically, the MSE of DS in (7) can be written as follows

$$\mathbb{MSE}_{\mathrm{DS}} \leq \mathrm{const} \cdot \lambda_p^2 \cdot (\sigma^2 + \mathbb{MSE}_{\mathrm{OLS}}).$$

In other words, even though no knowledge of the sparsity scenario was used in estimating $\hat{\theta}$, the estimation error is proportional to $\log p$ times the error rate achieved if the significant X's were known. So the price we pay for choosing the true predictors by DS is quite small as $\log p$ is not a fast rate. This feature is known as the oracle property of DS.

2.3.2 Dantzig Selector for Copula Score

It is not difficult to see that under Gaussian errors the constraint in (7) is a constraint on the score function of the underlying likelihood. So the DS can be equivalently interpreted as looking for a sparse θ close to the peak of the normal likelihood. This observation motivates the estimator we propose.

The Dantzig Selector SMLE (DS-SMLE) we propose is the solution to the following minimization problem

$$\min_{\beta,\gamma_N} ||\gamma_N||_{l_1} \quad \text{subject to} \quad \left|\left| \frac{1}{N} \sum_{i=1}^{N} \nabla_{(\beta,\gamma)} \ln c_{\gamma_N} \left(F(y_{i1};\beta), F(y_{i2};\beta)\right) \right|\right|_{l_\infty} \leq \ r \tag{8}$$

$$\text{and} \quad \frac{1}{N} \sum_{i=1}^{N} \nabla_\beta \ln f(y_{it};\beta) = \ 0, \ t = 1,2$$

where $c_\gamma(\mathbf{u}) = b_{\mathbf{k}}(\mathbf{u})$ is the Bernstein copula density, and $\nabla_{(\beta,\gamma)}$ denotes the derivative with respect to (β, γ).

The mean zero conditions on the marginal scores correspond to the assumption of correct specification of the marginals, which is our basic supposition. The copula score with respect to β and γ corresponds to the additional terms in the joint log-likelihood. In the fully parametric setting with a correctly-specified (up to a finite dimensional parameter γ) copula family, this score would be zero mean. In our setting, γ represents a function and $\dim\{\gamma\}$ is potentially greater than the sample size. Essentially, our estimator looks for such a vector (β', γ') for which γ has the smallest l_1-norm and the first order conditions characterizing the MLE solution hold within a fixed tolerance level.

This problem is an example of l_1-norm minimization subject to nonlinear constraints. There

are equivalent convex formulations for such problems (see, e.g., Candes, 2006). We can rewrite (8) as follows

$$\min_{\beta,\gamma_N,x} \sum_{j=1}^{\dim \gamma_N} x_i \quad \text{subject to} \quad \begin{array}{c} -x \preceq \gamma_N \preceq x \\ -r\mathbf{1} \preceq \frac{1}{N}\sum_{i=1}^{N} \nabla_{(\beta,\gamma)} \ln c_{\gamma_N}(F(y_{i1};\beta), F(y_{i2};\beta)) \preceq r\mathbf{1} \\ \frac{1}{N}\sum_{i=1}^{N} \nabla_{\beta} \ln f(y_{it};\beta) = \ 0, \ t = 1,2 \end{array} \tag{9}$$

where $x = \{x_i\}_{i=1}^{\dim \gamma_N}$, $\mathbf{1}$ denotes a conforming vector of ones and "$\preceq$" represents coordinate-wise comparison of vectors. This will be the preferred formulation in practice because standard convex optimization procedures and fast algorithms are available to compute the solution, which includes $\hat{\beta}$ (see, e.g., Birge and Massart, 1997; Devroye and Lugosi, 2000).

In order to see the relationship between this estimator and the penalized LS problem (6), note that DS-SMLE can be viewed as a solution to the following penalized MLE problem:

$$\min_{\beta,\gamma} \left\{ -\frac{1}{N} \ln L_{\gamma}(\beta) + r \sum_{j=1}^{\dim\{\gamma\}} |\gamma_j| \right\}, \tag{10}$$

where $\ln L_{\gamma}(\beta)$ is the copula-based log-likelihood given in (2), in which the marginals are assumed to be correctly specified. This is, of course, the penalized LS criterion from (6), with SSE replaced by $\ln L$, and the logic of our estimator is in essence the same as that of the conventional Dantzig selector – we are choosing the sparsest vector satisfying the Dantzig constraint implied by the penalized problem.

The choice of $\frac{1}{N} \ln L_{\gamma}(\beta)$ in (10) is natural if we view our problem as a minimization of the Kullback-Leibler distance between the true density $h(y_1, y_2)$ and the sieve-based density $h_{\gamma}(y_1, y_2; \beta)$, where $h_{\gamma}(y_1, y_2; \beta) = f(y_1; \beta) \cdot f(y_2; \beta) \cdot c_{\gamma}(F(y_1; \beta), F(y_2; \beta))$. Let $\mathbb{KL}(f, g)$ denote the Kullback-Leibler distance between arbitrary densities f and g. Then,

$$\arg\min_{\beta,\gamma} \mathbb{KL}(h, h_{\gamma}) = \arg\min_{\beta,\gamma} \mathbb{E} \ln \frac{h(y_1, y_2)}{h_{\gamma}(y_1, y_2; \beta)} = \arg\min_{\beta,\gamma} \left[-\mathbb{E} \ln h_{\gamma}(y_1, y_2; \beta) \right].$$

The expectation we minimize depends on the unknown h, so instead, we approximate it by its empirical counterpart $-\frac{1}{N} \ln L_{\gamma}(\beta)$. From this perspective, the problem in (10) can be viewed as a minimization of penalized Kullback-Leibler divergence.

2.3.3 Oracle Inequality

In this section we provide an oracle property of our estimator. We compare its risk with that of an infeasible procedure in which an oracle tells us which components of γ are insignificant. We start with a result for the copula parameter γ.

Suppose the marginal distributions are known. Then, the DS problem in (8) reduces to looking for the sparsest vector γ such that $\left|\left|\frac{1}{N}\sum_{i=1}^{N}\nabla_\gamma \ln c_\gamma\left(u_{i1},u_{i2}\right)\right|\right|_{l_\infty} \leq r$, where $u_{ij}=F(y_{ij}), j=1,2$, are obtained using the known marginals. Let $\hat{\gamma}$ denote this solution. The first result gives a bound on the KL divergence of the $\hat{\gamma}$-based copula.

Proposition 1 *Let $c_\gamma(\mathbf{u})$ be the Bernstein copula sieve, i.e. $c_\gamma(\mathbf{u})=\gamma'\mathbf{g}(\mathbf{u})$, where $\mathbf{g}(\mathbf{u})=(g_1(\mathbf{u}),\ldots,g_J(\mathbf{u}))'$ and $g_j(\mathbf{u})=\prod_{s=1}^{2}\beta(u_s;j_s,k_s-j_s+1), j=1,\ldots,J$. Let $M_j\equiv||g_j(\mathbf{u})||_{l_\infty}$, $j=1,\ldots,J$. Then, with probability close to one, for all $\gamma\in\mathbb{R}^J$*

$$\mathbb{KL}(c,c_{\hat{\gamma}})-2r\sum_{j=1}^{J}|\hat{\gamma}_j-\gamma_j|\leq\mathbb{KL}(c,c_\gamma) \tag{11}$$

Sketch of proof. Let $l_{\gamma i}\equiv\ln c_\gamma(u_{1i},u_{2i})$ and let $J\equiv\dim\{\gamma\}$. By definition of $\hat{\gamma}$,

$$-\frac{1}{N}\sum_{i=1}^{N}l_{\hat{\gamma}i}+r\sum_{j=1}^{J}|\hat{\gamma}_j|\leq-\frac{1}{N}\sum_{i=1}^{N}l_{\gamma i}+r\sum_{j=1}^{J}|\gamma_j|,$$

for any $\gamma\in\mathbb{R}^J$. Thus,

$$\mathbb{KL}(c,c_{\hat{\gamma}})\leq\mathbb{KL}(c,c_\gamma)+\frac{1}{N}\sum_{i=1}^{N}(l_{\hat{\gamma}i}-l_{\gamma i})-\mathbb{E}(l_{\hat{\gamma}i}-l_{\gamma i})+r\sum_{j=1}^{J}|\gamma_j|-r\sum_{j=1}^{J}|\hat{\gamma}_j|$$

Define $\xi_j(\mathbf{u_i})=\frac{g_j(\mathbf{u}_i)}{c_\gamma(\mathbf{u}_i)}$ and let $D_j=\frac{1}{N}\sum_{i=1}^{N}\{\xi_j(\mathbf{u}_i)-\mathbb{E}\xi_j(\mathbf{u}_i)\}$. Define the even $\Omega=\bigcap_{j=1}^{J}\{|D_j|\leq r\}$. By concavity of the log-function,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}(l_{\hat{\gamma}i}-l_{\gamma i})-\mathbb{E}(l_{\hat{\gamma}i}-l_{\gamma i}) &\leq \frac{1}{N}\sum_{i=1}^{N}\frac{1}{c_\gamma(\mathbf{u}_i)}\left[c_{\hat{\gamma}}(\mathbf{u}_i)-c_\gamma(\mathbf{u}_i)\right]-\mathbb{E}\frac{1}{c_\gamma(\mathbf{u}_i)}\left[c_{\hat{\gamma}}(\mathbf{u}_i)-c_\gamma(\mathbf{u}_i)\right]\\ &= \sum_{j=1}^{J}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{g_j(\mathbf{u}_i)}{c_\gamma(\mathbf{u}_i)}-\mathbb{E}\frac{g_j(\mathbf{u}_i)}{c_\gamma(\mathbf{u}_i)}\right)\left[\hat{\gamma}_j-\gamma_j\right]\end{aligned}$$

Therefore,

$$\mathbb{KL}(c, c_{\hat{\gamma}}) \leq \mathbb{KL}(c, c_{\gamma}) + \sum_{j=1}^{J} \left(\frac{1}{N} \sum_{i=1}^{N} \xi_j(\mathbf{u}_i) - \mathbb{E}\xi_j(\mathbf{u}_i) \right) [\hat{\gamma}_j - \gamma_j] + r \sum_{j=1}^{J} |\gamma_j| - r \sum_{j=1}^{J} |\hat{\gamma}_j|$$

Hence, on the event Ω,

$$\begin{aligned} \mathbb{KL}(c, c_{\hat{\gamma}}) &\leq \mathbb{KL}(c, c_{\gamma}) + r \sum_{j=1}^{J} |\hat{\gamma}_j - \gamma_j| + r \sum_{j=1}^{J} |\gamma_j| - r \sum_{j=1}^{J} |\hat{\gamma}_j| \\ &\leq 2r \sum_{j=1}^{J} |\hat{\gamma}_j - \gamma_{oj}|, \end{aligned}$$

where the last inequality follows by the triangle inequality.

Now by the Hoeffding inequality,

$$\mathbb{P}(\Omega) \leq \sum_{j=1}^{J} \mathbb{P}(|D_j| < r) \leq \sum_{j=1}^{J} \exp(nr^2/(16M_j^2)) = \delta.$$

3 Simulations

In this section we study the finite sample behavior of DS-SMLE as well as discuss issues arising when simulating from the Bernstein copula. Our goal is to compare the behavior of DS-SMLE with QMLE and SMLE, where the QMLE is the conventional estimator based on the independence assumption and the SMLE is the unpenalized SMLE based on the Bernstein copula. The DS-SMLE reduces to SMLE when $r = 0$.

Numerically, the fundamental difference between SMLE and DS-SMLE is that the SMLE estimates the entire vector γ_N for some large value of J_N, while DS-SMLE shrinks the elements of γ_N toward zero and estimates only the non-zero elements.

3.1 Simulating from Bernstein copula

A key issue in simulations is how to generate data from the Bernstein copula. The problem is that the standard way of generating observations from an arbitrary copula, known as the conditional cdf method, is too expensive in the settings of the Bernstein copula. The reason for this is that γ is obtained as the first order difference of parameters in the Bernstein copula cdf. As a result, γ basically contains $\Delta P_c(\frac{j_1}{k_1}, \frac{j_2}{k_2})$ and we have to solve a large system of equations to obtain $P_c(\frac{j_1}{k_1}, \frac{j_2}{k_2})$,

where

$$\Delta P_c\left(\frac{j_1}{k_1},\frac{j_2}{k_2}\right) = P_c\left(\frac{j_1+1}{k_1},\frac{j_2+1}{k_2}\right) - P_c\left(\frac{j_1+1}{k_1},\frac{j_2}{k_2}\right) - P_c\left(\frac{j_1}{k_1},\frac{j_2+1}{k_2}\right) + P_c\left(\frac{j_1}{k_1},\frac{j_2}{k_2}\right)$$

As an alternative, we use the accept-reject approach (see, e.g., Pfeifer, Strassburger, and Philipps, 2009). To introduce the method, suppose we want to generate data from a distribution F with a pdf $f(x)$, which is a complicated distribution and we do not know how to simulate from it directly. The basic idea of the method is to find another distribution G with a pdf $g(y)$, for which we already have an efficient algorithm to generate data. The key is that this distribution should also be very close to $f(x)$. Specifically, the ratio $f(x)/g(x)$ should be bounded by a positive constant M, i.e. $\sup_x\{f(x)/g(x)\} \leq M$. The we can apply the following procedure:

1. Generate y from $g(y)$
2. Independently generate u from uniform on [0,1]
3. If $u \leq \frac{f(y)}{Mg(y)}$, then set $x = y$ and use x as a sample from $f(x)$. Otherwise, go back to Step 1.

It can be easily shown that $P(Y \leq y|U \leq \frac{f(y)}{cg(y)}) = F(y)$. Also, note that the expected number of steps required to generate one observation from $f(x)$ is M.

We wish to apply the accept-reject method to the Bernstein copula. We use a multivariate uniform distribution as the reference distribution $G(.)$ with the density function $g(.) = 1$. In this case, $M = \sup_u\{b_{\mathbf{k}}(\mathbf{u})/g(\mathbf{u})\} = \max_{\mathbf{u}\in[0,1]^d}\{b_{\mathbf{k}}(\mathbf{u})\}$. The simulation algorithm is as follows:

1. Generate $(u_1,\ldots,u_d)$ from the multivariate uniform distribution. Here d denotes the number of cross-sections.
2. Independently generate u_{d+1} from uniform on [0,1].
3. if $u_{d+1} \leq \frac{b_{\mathbf{k}}(\mathbf{u})}{M}$, then use $(u_1,\ldots,u_d)$ as an observation from the Bernstein copula. Otherwise, go back to step 1.

It is clear that due to the reference distribution G being uniform, we can actually combine Step 1 and 2 into one step.

3.2 Sparse parameter path

The tuning parameter r is key to the amount of shrinkage done by the DS. As a first step of the simulation exersice we study the behavior of our estimator of γ over all r.

Our data generating process has exponential marginals with $\mu_1 = \mu_2 = 0.5$ and the Bernstein copula with $J = 25$ (five parameters in each dimension), so in total, there are 27 parameters. However, the γ has only five elements out 25 that are nonzero as shown in the following matrix

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0.228 \\ 0 & 0 & 0 & 0.141 & 0 \\ 0 & 0 & 0.262 & 0 & 0 \\ 0 & 0.141 & 0 & 0 & 0 \\ 0.228 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

This corresponds to a copula with a high negative dependence. The number of observations is 1000.

Figure 1 shows the estimated parameter paths for the non-zero elements of γ (colored solid lines) and the insignificant elements (dashed red lines). There are two important observations. First, the DSSMLE can correctly identify the non-zero elements in γ. Second, in the region where the zero γ_j's are actually estimated to be close to zero (the region with small l_1), the non-zero γ_j's are estimate to be smaller than the true values. This suggests that the DS-SMLE over-shrinks γ.

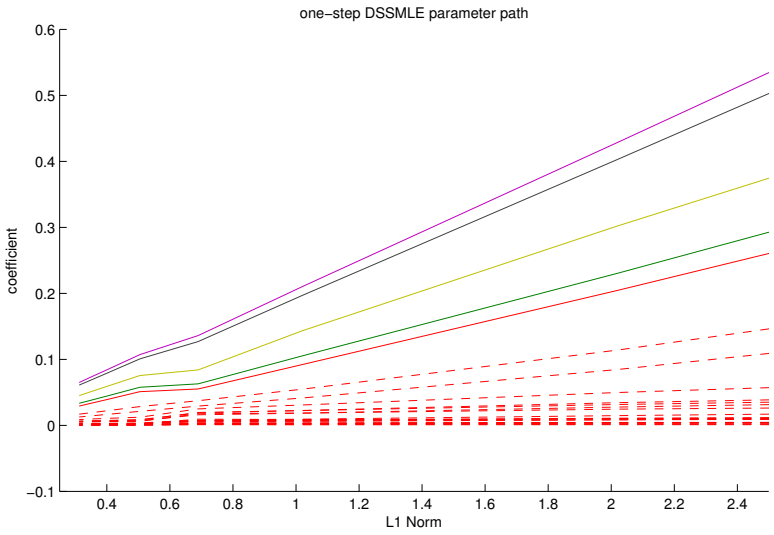
The over-shrinkage result is not uncommon in the DS literature and James and Radchenko (2009) propose a two-step procedure called double Dantzig to overcome this issue. We follow James and Radchenko (2009) and implement the following two-step procedure in our simulations:

1. Run the DS-SMLE using a large value of the tuning parameter. Select the non-zero elements γ_j. Denote the selected set by γ^*.

2. Run the unrestricted SMLE over γ^* and β.

So in effect we run two DS-SMLE where in the second step we set the tuning parameter equal to be zero. A similar procedure called the gaussian Dantzig selector was proposed by (Candes and Tao, 2007, p. 2323) and can be seen as a special case of the double Dantzig of James and Radchenko (2009).

Figure 1: DSSMLE Parameter Path



Notes: Plot of estimated coefficients for different values of λ. The solid lines represent the variables which are nonzeros in the true setting of γ. The dashed lines correspond to the remaining variables.

3.3 Simulation results

Compared to the QML and SMLE, our DS-SMLE stimator does not restrict the dependence structure but uses a sparsity scenario, that is, it estimates only non-zero elements of γ. For all three estimators, we report bias, variance, MSE, relative effiency (RE) with respect to the QMLE and relative MSE (RMSE) with respect to QMLE. For SMLE and DS-SMLE we also report the dimension of γ. The number of observations is 500 and the number of replications is 1,000.

We consider three data generation processes. All have the same exponential marginals, where the mean μ is the parameter of interest with the true value $\mu_1 = \mu_2 = 0.5$, but the copula functions are different. We use the Plackett, Student-t, and Frank copulas as these copula families are often encountered in simulations. The copula parameter varies over the relevant range, representing different strengths of dependence. We report Kendall's τ for each such value.

Table 1-8 summarize the results. Two things are important here. First, for some values of τ, the DS-SMLE is at least as efficient as unrestricted SMLE, while it dramatically reduces the number of sieve parameters to be estimated. For example, Table 1 shows that the DS-SMLE estimates only 15 of 256 sieve parameters and it preserves the efficiency gains of the SMLE. Second, as negative dependence goes from high to low, both the SMLE and DS-SMLE have decreasing relative efficiency over QMLE. For instance, in Table 1-3, we can observe this for the Plackett copula, and similarly

in Table 4-6 for t and Table 7-8 for Frank.

4 Application from Insurance

We illustrate the use of the DS-SMLE with an insurance application. We consider automobile bodily injury liability claims from a sample of $n = 29$ Massachusetts towns in 1995 and 1997. The details of the data set can be found in Frees and Wang (2005). The two cross-sections have a strong positive correlation at 0.88 in the average town-wide claims (AC).

Following Frees and Wang (2005), the claims are assumed to have the same gamma distribution for the two years and the goal is the efficient estimation of the parameters (a, b) of that distribution. That is, we use the following cdf and pdf, respectively:

$$\begin{aligned} F_i(x|a,b) &= \frac{1}{b^a\Gamma(a)}\int_0^x t^{a-1}e^{\frac{-t}{b}}dt \\ f_i(x|a,b) &= \frac{1}{b^a\Gamma(a)}x^{a-1}e^{\frac{-x}{b}} \qquad \text{where i=1,2} \end{aligned}$$

The four estimators we consider are QMLE, Pseudo-MLE (PMLE), SMLE, and DS-SMLE. The QMLE estimator assumes independence between cross-section. It is known to be consistent even if the independence assumption is incorrect. To obtain a robust estimator of the standard errors, the "sandwich" formula is used. The PMLE is the estimator based on a fully specified parametric joint likelihood. We follow Frees and Wang (2005) and use t-copula for this. The PMLE is consistent if the assume copula family is correct. Otherwise, the PMLE is generally biased and we do not know either the sign or the magnitude of the bias. Both the SMLE and DS-SMLE are robust in the sense that they do not depend on a specific assumption on the copula family. They are more efficient asymptotically relative to QMLE and, as illustrated by the simulation of the previous sections, behavie similarly in small samples.

Table 9 report the estimates and standard errors. A few interesting observations can be made using these results. First, both the SMLE and DS-SMLE have smaller standard errors than QMLE. Second, while the SMLE shows evidence of bias, the DS-SMLE estimates are fairly close to FMLE or QMLE. We use 8 parameters in each dimension of the sieve, where this value is chosen using the BIC criterion. So for the SMLE, we have 66 parameter to estimate. For the DS-SMLE, we have only 10.

5 Concluding remarks

We have proposed to use a penalized sieve to improve efficiency of likelihood-based estimators in panel settings. The settings can be easily generalized to multivariate models where a part of the joint distribution is modelled by a sieve with a potentially very large number of parameters, only a few of which are significant.

We showed that the sparse sieve MLE, based on the Dantzig penalization, has very similar properties to the sieve MLE in finite samples, so the sparsity imposed by the Dantzig constraint does not add to the bias as much as much as it takes away from the variance. We also looked at the behavior of the estimator for various values of the tolerance and found evidence that our estimator tends to over-shrink. We proposes a two-step procedure that addresses this issue and clarifies the problem of choosing the tolerance level.

The relative efficiency and mean square gains we obtain are up to 70% which is very encouraging. The computational benefit is of course even more important; especially in cases when SMLE is infeasible due to small sample size.

Table 1: Comparison of QMLE, SMLE and DS-SMLE with Kendall $\tau = -0.9$, Plackett $\theta = 0.002$, $dim(\gamma_{smle}) = 256$,$dim(\gamma_{ds-smle} = 15)$

$\tau = -0.9$	μ_1 QMLE	SMLE	DS-SMLE	μ_2 QMLE	SMLE	DS-SMLE
mean	0.4996	0.4939	0.4959	0.4999	0.4940	0.4956
Nvar	0.2575	0.0769	0.0735	0.2864	0.0771	0.0750
MSE	0.2574	0.1137	0.0901	0.2862	0.1124	0.0945
RE		0.2988	0.2855		0.2692	0.2619
RMSE		0.4419	0.3499		0.3929	0.3303

Table 2: Comparison of QMLE, SMLE and DS-SMLE with Kendall $\tau = -0.8$ Plackett $\theta = 0.009$,$dim(\gamma_{smle}) = 49$,$dim(\gamma_{ds-smle} = 8)$

$\tau = -0.8$	μ_1 QMLE	SMLE	DS-SMLE	μ_2 QMLE	SMLE	DS-SMLE
mean	0.4999	0.4903	0.4941	0.5003	0.4898	0.4937
Nvar	0.4836	0.3491	0.2888	0.4976	0.3376	0.2951
MSE	0.4831	0.4433	0.3228	0.4971	0.4405	0.3340
RE		0.7220	0.5972		0.6785	0.5932
RMSE		0.9177	0.6682		0.8862	0.6718

Table 3: Comparison of QMLE, SMLE and DS-SMLE with Kendall $\tau = -0.7$ Plackett $\theta = 0.023$,$dim(\gamma_{smle}) = 64$,$dim(\gamma_{ds-smle} = 61)$

$\tau = -0.7$	μ_1 QMLE	SMLE	DS-SMLE	μ_2 QMLE	SMLE	DS-SMLE
mean	0.5005	0.4947	0.4951	0.4995	0.4942	0.4936
Nvar	0.2455	0.2013	0.2021	0.2508	0.2157	0.2099
MSE	0.2451	0.2148	0.2137	0.2504	0.2323	0.2296
RE		0.8198	0.8233		0.8598	0.8367
RMSE		0.8764	0.8717		0.9277	0.9170

Table 4: Comparison of QMLE, SMLE and DS-SMLE with Kendall $\tau = -0.9$
t copula,$R = -0.9877, \mu = 20$, $dim(\gamma_{smle}) = 225$,$dim(\gamma_{ds-smle} = 150)$

$\tau = -0.9$	μ_1 QMLE	SMLE	DS-SMLE	μ_2 QMLE	SMLE	DS-SMLE
mean	0.5016	0.4941	0.4939	0.4993	0.4932	0.4935
var	0.4983	0.1404	0.1561	0.4933	0.1470	0.1782
MSE	0.5003	0.1747	0.1933	0.4934	0.1932	0.2209
RE		0.2817	0.3132		0.2980	0.3612
RMSE		0.3491	0.3864		0.3915	0.4477

Table 5: Comparison of QMLE, SMLE and DS-SMLE with Kendall $\tau = -0.8$
t copula,$R = -0.9511, \mu = 20$,$dim(\gamma_{smle}) = 225$,$dim(\gamma_{ds-smle} = 64)$

$\tau = -0.8$	μ_1 QMLE	SMLE	DS-SMLE	μ_2 QMLE	SMLE	DS-SMLE
mean	0.4990	0.4933	0.4927	0.5006	0.4940	0.4935
var	0.4889	0.2904	0.2430	0.5023	0.2965	0.2411
MSE	0.4895	0.3356	0.2967	0.5021	0.3327	0.2835
RE		0.5940	0.4971		0.5904	0.4801
RMSE		0.6857	0.6062		0.6626	0.5647

Table 6: Comparison of QMLE, SMLE and DS-SMLE with Kendall $\tau = -0.7$
t copula, $R = -0.8910, \mu = 20$,$dim(\gamma_{smle}) = 64$,$dim(\gamma_{ds-smle} = 62)$

$\tau = -0.7$	μ_1 QMLE	SMLE	DS-SMLE	μ_2 QMLE	SMLE	DS-SMLE
mean	0.4991	0.4945	0.4947	0.5006	0.4953	0.4955
var	0.4953	0.4377	0.4017	0.4944	0.4281	0.4069
MSE	0.4956	0.4671	0.4290	0.4943	0.4497	0.4269
RE		0.8836	0.8110		0.8659	0.8230
RMSE		0.9425	0.8656		0.9099	0.8638

Table 7: Comparison of QMLE, SMLE and DS-SMLE with Kendall $\tau = -0.9$
Frank $\theta = -39$,$dim(\gamma_{smle}) = 225$,$dim(\gamma_{ds-smle} = 111)$

$\tau = -0.9$	μ_1 QMLE	SMLE	DS-SMLE	μ_2 QMLE	SMLE	DS-SMLE
mean	0.4996	0.4926	0.4927	0.5006	0.4930	0.4929
Nvar	0.4772	0.1479	0.1552	0.4823	0.1477	0.1547
MSE	0.4769	0.2032	0.2077	0.4821	0.1968	0.2044
RE		0.3099	0.3252		0.3062	0.3208
RMSE		0.4262	0.4355		0.4083	0.4241

Table 8: Comparison of QMLE, SMLE and DS-SMLE with Kendall $\tau = -0.8$
Frank $\theta = -18.19$,$dim(\gamma_{smle}) = 49$,$dim(\gamma_{ds-smle} = 7)$

$\tau = -0.8$	μ_1 QMLE	SMLE	DS-SMLE	μ_2 QMLE	SMLE	DS-SMLE
mean	0.4996	0.4879	0.4947	0.5004	0.4875	0.4886
Nvar	0.4790	0.3131	0.2787	0.4899	0.3139	0.2802
MSE	0.4787	0.4580	0.3062	0.4896	0.4708	0.4098
RE		0.6538	0.5819		0.6408	0.5720
RMSE		0.9569	0.6397		0.9616	0.8371

Table 9: QMLE, t copula based Pseudo-MLE, SMLE, DS-SMLE for insurance application with standard errors

	QMLE	PMLE	SMLE		DS-SMLE	
	(Rob.St.Er)	(Rob.St.Er)	(St.Er	Rob.St.Er)	(St.Er	Rob.St.Er)
a	14.7561	15.0103	15.7039		15.0344	
	(4.4702)	(4.3306)	(3.1607	6.1343)	(3.4653	6.0796)
b	9.7020	9.6806	9.2871		9.7482	
	(2.9080)	(2.8499)	(2.1433	3.4674)	(2.4846	3.8158)
LogL	-290.8190	-266.3389	-271.5390		-271.7004	

References

Amsler, C., A. Prokhorov, and P. Schmidt (2013): "Using Copulas to Model Time Dependence in Stochastic Frontier Models," *Econometric Reviews - forthcoming*.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Statistics*, 37(4), 1705–1732.

Birge, L., and P. Massart (1997): "From model selection to adaptive estimation," in *Festschrift for Lucien LeCam: Research Papers in Probability and Statistics*, ed. by D. Pollard, E. Torgersen, and C. Yang, pp. 55–87.

Bouezmarni, T., and J. V. K. Rombouts (2010): "Nonparametric density estimation for multivariate bounded data," *Journal of Statistical Planning and Inference*, 140(1), 139–152.

Bunea, F., A. B. Tsybakov, M. H. Wegkamp, and A. Barbu (2010): "Spades and mixture models," *The Annals of Statistics*, 38(4), 2525–2558.

Burda, M., and A. Prokhorov (2013): "Copula-Based Factorization for Bayesian Infinite Mixture Models," *Concordia University Working paper*.

Candes, E., and T. Tao (2007): "The Dantzig Selector: Statistical Estimation When p Is Much Larger than n," *The Annals of Statistics*, 35(6), pp. 2313–2351.

Candes, E. J. (2006): "Modern statistical estimation via oracle inequalities," *Acta Numerica*, 15, 257–325.

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, pp. 5549–5632.

Devroye, L., and G. Lugosi (2000): *Combinatorial Methods in Density Estimation*. Springer.

Frees, E. W., and P. Wang (2005): "Credibility using copulas," *North American Actuarial Journal*, 9(2), 31–48.

James, G. M., and P. Radchenko (2009): "A generalized Dantzig selector with shrinkage tuning," *Biometrika*, 96(2), 323–337.

Koltchinskii, V. (2009): "The Dantzig selector and sparsity oracle inequalities," *Bernoulli*, 15(3), 799–828.

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

Panchenko, V., and A. Prokhorov (2013): "Efficient Estimation of Parameters in Marginals," *Concordia University Working Paper*.

Pfeifer, D., D. Strassburger, and J. Philipps (2009): "Modelling and simulation of dependence structures in nonlife insurance with Bernstein copulas," in *39th International ASTIN Colloquium, Helsinki*.

Prokhorov, A., and P. Schmidt (2009): "Likelihood-based estimation in a panel setting: robustness, redundancy and validity of copulas," *Journal of Econometrics*, 153(1), 93–104.

Sancetta, A., and S. Satchell (2004): "The Bernstein Copula And Its Applications To Modeling And Approximations Of Multivariate Distributions," *Econometric Theory*, 20(03), 535–562.

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(2), 580–615.

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

Tenbusch, A. (1994): "Two-dimensional Bernstein polynomial density estimators," *Metrika*, 41(1), 233–253, Metrika.

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

Zheng, Y. (2011): "Shape restriction of the multi-dimensional Bernstein prior for density functions," *Statistics and Probability Letters*, 81(6), 647–651.