C.I.R.P.É.E

Centre Interuniversitaire sur le Risque,
les Politiques Économiques et l'Emploi

Cahier de recherche/Working Paper **07-31**

# Semiparametric Multivariate Density Estimation for Positive Data Using Copulas

Taoufik Bouezmarni
Jeroen V.K. Rombouts

Octobre/October 2007

Bouezmarni : HEC Montréal, CREF, Institute of Statistics (Université Catholique de Louvain)
Rombouts : Institute of Applied Economics at HEC Montréal, CIRANO, CIRPÉE, CREF 3000, chemin de la Côte-Sainte-Catherine, Montréal (Québec) Canada H3T 2A7

Financial support from the Centre for research on e-finance at HEC Montréal is greatly acknowledged.

**Abstract:** In this paper we estimate density functions for positive multivariate data. We propose a semiparametric approach. The estimator combines gamma kernels or local linear kernels, also called boundary kernels, for the estimation of the marginal densities with semiparametric copulas to model the dependence. This semiparametric approach is robust both to the well known boundary bias problem and the curse of dimensionality problem. We derive the mean integrated squared error properties, including the rate of convergence, the uniform strong consistency and the asymptotic normality. A simulation study investigates the finite sample performance of the estimator. We find that univariate least squares cross validation, to choose the bandwidth for the estimation of the marginal densities, works well and that the estimator we propose performs very well also for data with unbounded support. Applications in the field of finance are provided.

**Keywords:** Asymptotic properties, asymmetric kernels, boundary bias, copula, curse of dimension, least squares cross validation

**JEL Classification:** C13, C14, C22

**Résumé:** Dans cet article nous estimons la fonction de densité pour des données multivariées et positives. Nous proposons une approche semi-paramétrique. La méthode utilise l'estimateur à noyau gamma ou local linéaire pour évaluer les densités marginales et la copule semi-paramétrique pour modéliser la dépendance. Cette approche semi-paramétrique est robuste à la fois au problème de biais à la frontière et au problème de dimensionnalité. Nous dérivons l'erreur quadratique moyenne intégrée, y compris le taux de convergence, la convergence uniforme presque sûre ainsi que la normalité asymptotique. Une étude Monte Carlo montre la performance de cet estimateur. Pour choisir le paramètre de lissage on propose d'utiliser la méthode de validation croisée des moindres carrés. Nous montrerons par simlations la performance de cette technique. Des applications dans le domaine des finances sont fournies.

**Mots Clés :** Propriétés asymptotiques, noyaux asymétriques, problème du biais, copule, problème de dimension, validation croisée.

# 1 Introduction

Many results on nonparametric density estimation are based on the assumption that the support of the random variable of interest is the real line. However, in applications, data are often bounded with a possible high concentration close to the boundary. For example, in labor economics, the income distribution for a specific country is bounded at the minimum wage. Usual nonparametric density estimation techniques, for example the well known Gaussian kernel, for these kind of data produce inconsistent results because the kernel allocates weight outside the support implying an underestimation of the underlying density in the boundary. This boundary bias problem is well documented in the univariate case. The first technique to resolve this problem is proposed by Schuster (1985) suggesting the reflection method. Lejeune and Sarda (1992), Jones (1993) Jones and Foster (1996), Müller (1991), and Rice (1984) use flexible kernels called boundary kernels instead of the usual fixed kernels. Marron and Ruppert (1994) recommend to transform data before applying the standard kernel. Chen (2000) proposes a gamma kernel estimator, Bouezmarni and Scaillet (2005) and Bouezmarni and Rombouts (2006) investigate the properties of a gamma estimator in respectively a mean absolute deviation and a time series framework.

In general, the univariate framework is only a first step towards multivariate density estimation in order to explain links between variables the supports of some are potentially bounded. The problem of inconsistent density estimation carries over (and becomes even more substantial) in the case of multivariate bounded random variables. For the same reason as above, the multivariate Gaussian kernel density estimator is not suitable for these kind of random variables. An additional problem with nonparametric multivariate density estimation is that the rate of convergence of the mean integrated squared error increases with the dimension. This is the well known curse of dimensionality problem. To date, the boundary and the curse of dimension problems have not been addressed simultaneously. For example, Müller and Stadtmüller (1999) propose a multivariate estimator without a boundary problem but with a problem of curse of dimension. Liebscher (2005) puts forward a semiparametric estimator based on copulas and on the standard kernel estimator for the marginal densities which solves the curse of dimension problem but not the boundary problem.

This paper proposes a multivariate semiparametric density estimation method which is robust to both the boundary and the curse of dimension problem. The estimator combines gamma or local linear kernels the support of which matches that one of the underlying multivariate density,

and semiparametric copulas. This leads to an estimator which is easy to implement. We derive asymptotic properties such as the mean integrated squared error, uniform strong consistency and asymptotic normality. In the simulations we compare the finite sample performance of the (modified) gamma and the local linear estimator for the marginal densities using the Gaussian and the Gumbel-Hougaard copula. We find that the univariate least squares cross validation technique to choose the bandwidths for the marginal kernel density estimators works successfully. Therefore, bandwidth selection for our estimator can be done in a computational straightforward manner. The simulations reveal also that for data without a boundary problem our estimator performs very well.

Examples of multivariate positive data abound in finance and economics. Cho (1998) investigates whether ownership structure affects investment using variables such as capital expenditures, and research and development expenditures sampled from the 1991 Fortune 500 manufacturing firms. Grullon and Michaely (2002) study the relationship over time between dividends and share repurchases conditional on the market value and the book value of assets for US corporations. In our application we estimate the joint density of the stock price and the total number of shares outstanding. The data come from 558 US companies observed in 2005. We test if the density depends on the fact that dividends are paid out or not, and on the fact that there is debt outstanding or not. We use the Gumbel-Hougaard copula as suggested by the simulation results.

The paper is organized as follows. The semiparametric estimator for multivariate positive data is introduced in Section 2. Section 3 provides convergence properties. In Section 4, we investigate the finite sample properties of the gamma and local linear kernel semiparametric copula estimator for positive bivariate data. Section 5 contains the application described above. Section 6 concludes. The proofs of the asymptotic results are gathered in the appendix.

## 2 Semiparametric density estimator

Let $X = \{(X_i^1, ..., X_i^d), i = 1, .., n\}$ be a sample of independent and identically distributed random variables in $R^{+d}$, with distribution function $F$ and density function $f$. We estimate the density function with a semiparametric method based on nonparametric marginal density estimates and a semiparametric copula. Compared to a full nonparametric approach we impose some structure on the unknown distribution but doing so we do not have the curse of dimension problem. Furthermore,

in several research fields one wants to interpret parameters of interest that measure the association between the random variables. What is not of interest is left unspecified.

From Sklar (1959) it is well known that the distribution function can be expressed via a copula

$$F(x_1, ..., x_d) = \Gamma(F_1(x_1), ..., F_d(x_d)) \tag{1}$$

where $F_i$ is the marginal distribution of the random variable $X_i$, $\Gamma$ is a copula function which captures the dependence of $X$. See Nelson (1999) for a textbook reference on copulas. There are several possibilities to work with copulas. First, one can assume parametric models for both the copula and the marginal distribution. Estimation of the parameters is done by maximum likelihood or inference function for margins. See Oakes (1982), Romano (2002) and Joe (2005) for details of these methods. A second possibility is to consider nonparametric models for both the marginal distribution and the copula. Deheuvels (1979) proposes a method based on the multivariate empirical distribution. Gijbels and Mielniczuk (1990) use the kernel method to estimate a bivariate copula and suggest to use the reflection method to overcome the boundary bias problem. More recently, Chen and Huang (2007) propose a bivariate estimator based on the local linear estimator. A Bernstein polynomial kernel type estimator is developed by Sancetta and Satchell (2004) and Rödel (1987) uses the orthogonal series method. A third possibility to work with copulas is a semiparametric approach which supposes a parametric model for the copula, $\Gamma = \Gamma_\theta$, and a nonparametric model for the marginal distributions. This method is developed by Oakes (1986), and Genest, Ghoudi, and Rivest (1995) and Genest and Rivest (1993). Recently, Kim, Silvapulle, and Silvapulle (2007) compare semiparametric and parametric methods for estimating copulas.

In this paper our interest lies in the density function. It is well known that, by deriving (1) with respect to $(x_1, ..., x_d)$, the density function can be expressed as

$$f(x_1, ..., x_d) = f_1(x_1)...f_d(x_d)\gamma(F_1(x_1), ..., F_d(x_d)) \tag{2}$$

where $f_j$ is the marginal density of the random variable $X^j$ and $\gamma$ is the copula density. We estimate the density function in a semiparametric way. With respect to the semiparametric copula, we estimate the parameter $\theta$ by a consistent estimator. The distribution function of $X^j$ is estimated by $F_{nj}$ using the empirical distribution. The marginal density of $X^j = (X_1^j, ..., X_n^j)$ is estimated nonparametrically as

$$\hat{f}_j(x_j) = \frac{1}{n}\sum_{i=1}^{n} K(b_j, X_i^j)(x_j) \tag{3}$$

where $b_j$ is the bandwidth parameter and the kernel K is the local linear kernel when it is defined as

$$K_L(h,t)(x) = K_l\left(x, h, \frac{x-t}{h}\right) \tag{4}$$

where

$$K_l(x,h,t) = \frac{a_2(x,h) - a_1(x,h)t}{a_0(x,h)a_2(x,h) - a_1^2(x,h)} K(t), \tag{5}$$

$K$ is any symmetric kernel with a compact support $[-1,1]$ and

$$a_s(x,h) = \int_{-1}^{x/h} t^s K(t)dt. \tag{6}$$

We also consider a gamma kernel defined as

$$K_G(b,t)(x) = \frac{t^{x/b}\exp(-t/b)}{b^{x/b+1}\Gamma(x/b+1)} \tag{7}$$

and a modified gamma kernel

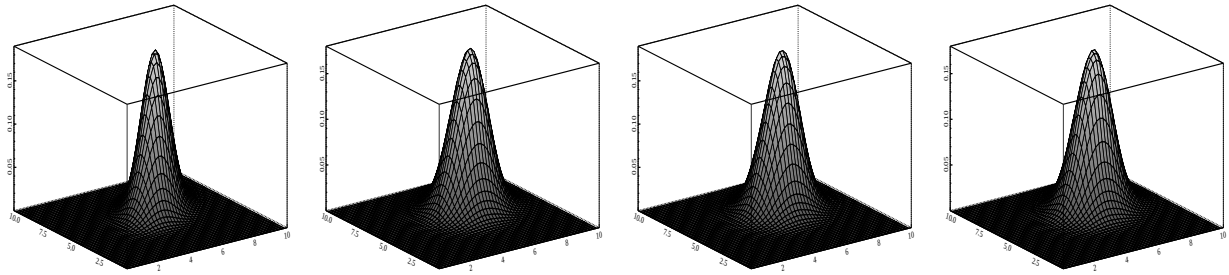
$$K_{MG}(b,t)(x) = \frac{t^{\rho(x)-1}\exp(-t/b)}{b^{\rho(x)}\Gamma(\rho(x))}, \tag{8}$$

where

$$\rho(x) = \begin{cases} x/b & \text{if} \quad x \geq 2b \\ \\ \frac{1}{4}(x/b)^2 + 1 & \text{if} \quad x \in [0, 2b). \end{cases} \tag{9}$$

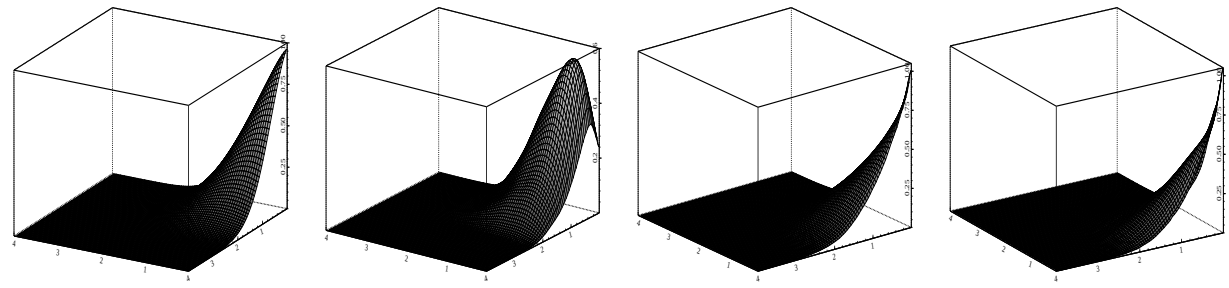
The second gamma kernel is proposed by Chen (2000) in order to reduce the bias of the gamma kernel $K_G$. In fact, in the next section we show that for this kernel the first derivative disappears in the asymptotic integrated bias.

To conclude, the semiparametric method separates the multivariate density estimator into marginal density estimation and copula estimation. With the univariate boundary kernels we resolve the potential boundary problem in the marginal densities, and the use of a semiparametric copula circumvents the curse of dimension problem. Therefore, to estimate the multivariate density we need to choose $n$ bandwidths and a copula family. Figure 1 displays shapes of the Gaussian, local linear and the gamma kernel estimator with a Gaussian copula for data without a boundary problem. We observe that the shapes of all the kernels are quite similar, demonstrating the flexibility of the local linear and the gamma kernels using a Gaussian copula. Figure 2 shows how the semiparametric estimator adapts nicely to densities with high a concentration in the boundary region and that the Gaussian kernel (panel b) is inconsistent for this type of data.



(a) Density (b) Gaussian kernel (c) Local linear kernel (d) Gamma kernel

Figure 1: Normal density function with Gaussian, local linear and gamma kernel estimators.



(a) Density (b) Gaussian kernel (c) Local linear kernel (d) Gamma kernel

Figure 2: Truncated normal with two boundary problems with Gaussian, local linear and gamma kernel estimators.

## 3 Convergence properties

In this section we establish the asymptotic properties of the semiparametric estimator. Assumptions on the bandwidth parameters and the copula parameter are given next.

**Assumptions on the bandwidth parameters**

**A1.** $a_j \to 0, \quad \text{and} \quad n^{-1}a_j^{-\frac{1}{2}} \to 0, \ \text{ for } j = 1, ..., d., \quad \text{as } n \to \infty.$

**A2.** $a_j \to 0, \quad \text{and} \quad \log(n)n^{-1}a_j^{-\frac{1}{2}} \to 0, \ \text{ for } j = 1, ..., d., \quad \text{as } n \to \infty.$

The condition A1 is needed for mean integrated squared error and the normality of the estimator, the condition A2 is required for the uniform strong convergence of the estimator. These conditions are similar to those of Bouezmarni and Scaillet (2005).

**Assumptions on the copula**

**P1.** Suppose that $\gamma_\theta$ is bounded on $[0,1]^d$ and

$$|\gamma_t(u_1, ..., u_d) - \gamma_s(v_1, ..., v_d)| \leq C \left( \sum_{i=1}^{d} |u_i - v_i| + |t - s| \right)$$

for $u = (u_1, ..., u_d), v \in J \subset [0,1]^d$, $t, s \in \Theta$, C is a constant and $J$ is the intersection of an open set and $[0,1]^d$.

**P2.**

$$||\hat{\theta} - \theta|| = O\left( \sqrt{\frac{ln(n)}{n}} \right), \quad \text{a.s.} \tag{10}$$

**P3.**

$$E(\hat{\theta} - \theta)^2 = O(\ln(n)n^{-1}). \tag{11}$$

The condition P1 allows to separate the two random terms, that is the parameter estimator and the marginal distribution estimators, in the copula estimator. Hence, it suffices to make assumptions P2 and P3 on the parameter estimator of the copula, since it is well known that the consistency of the empirical distribution estimator is guaranteed. Liebscher (2005) shows for the Raftery family and Gumbel family of copulas that the three conditions above are fulfilled.

Under the previous assumptions we establish our main theoretical results. The next proposition shows the asymptotic mean integrated squared error.

**Proposition 1. mean integrated squared error of $\hat{f}_{sp}$**

*Suppose that $f_1, ..., f_d$ are twice differentiable at $x$. Under assumption A1, P1 and P3 we have*

$$MISE = \int (\sum_{j=1}^{d} a_j B_j^*(x))^2 \, dx + \frac{1}{n} \left( \sum_{j=1}^{d} a_j^{-1/2} \int V_j(x) dx \right) + o\left( \sum_{j=1}^{d} a_j^2 \right) + o\left( n^{-1} \sum_{j=1}^{d} a_j^{-1/2} \right)$$

*where for the gamma kernel, $a_j = b_j$ and*

$$B_j = \gamma_\theta(x) \tilde{f}_j(x) B_j(x) \quad \textit{and} \quad V_j(x) = (2\sqrt{\pi})^{-1} \gamma_\theta^2(x) \tilde{f}_j^2(x) f_j(x_j) x_j^{-1/2}$$

*with*

$$\tilde{f}_j(x) = \prod_{k \neq j} f_k(x_k).$$

*The optimal bandwidths which minimize the asymptotic mean integrated squared error are*

$$a_j^* = c_j^* n^{-\frac{2}{5}}, \quad \textit{for some positive constants } c_1^*, ..., c_d^*. \tag{12}$$

*Therefore, the optimal asymptotic mean integrated squared error is*

$$AMISE^* = \left\{ \int (\sum_{j=1}^{d} C_j^* B_i^*(x))^2 \, dx + \left( \sum_{j=1}^{d} C_j^{*-1/2} \int V_j(x) \, dx \right) \right\} n^{-\frac{4}{5}} \tag{13}$$

■

In particular, if $a = a_1 = ... = a_d$, the optimal bandwidths and the optimal asymptotic mean integrated squared error are

$$a^* = \left( \frac{1}{4} \frac{\sum \int V_j(x) dx}{\sum \int B_j^*(x) dx} \right)^{\frac{2}{5}} n^{-\frac{2}{5}}, \quad \text{and} \quad AMISE^* = \frac{5}{4^{\frac{4}{5}}} \left( \sum \int V_j(x) dx \right)^{\frac{4}{5}} \left( \sum \int B_j^*(x) dx \right)^{\frac{1}{5}} n^{-\frac{4}{5}}$$

proposition 1 states the mean integrated squared error and the optimal bandwidth of the semi-parametric gamma estimator. The estimator is free from the curse of dimension since the rate of convergence is the same as in the univariate case. The optimal bandwidth can not be used in practice since it depends on the unknown density function. However, we can use for example least squares cross validation methods choosing optimal bandwidths for the marginal densities by noting that the same rate of convergence of mean integrated squared error for the multivariate estimator is obtained. The following remark states the MISE of the semiparametric estimator with the local linear estimator and the second gamma kernel estimator for the marginals.

**Remark 1.** *The results of proposition 1 remain valid*

- *For the local linear estimator with* $a_j = h_j^2$,

$$B_j^*(x) = \gamma_\theta(x)\tilde{f}_j(x)\frac{\kappa_2}{2}f^{jj}(x) \quad and \quad V_j(x) = \kappa^d\gamma_\theta(x)^2(x)\tilde{f}_j^2(x)f_j(x_j)$$

  *where* $\kappa_2 = \int x^2K(x)dx$ *and* $\kappa = \int K^2(x)dx$.

- *For the modified gamma kernel,* $a_j = b_j$,

$$B_j^*(x) = \gamma_\theta(x)\tilde{f}_j(x)\frac{x_j f^{jj}(x)}{2} \quad and \quad V_j(x) = (2\sqrt{\pi})^{-1}\gamma_\theta(x)^2(x)\tilde{f}_j^2(x)f_j(x_j)x_j^{-1/2}.$$

The following proposition establishes the uniform strong consistency of the semiparametric density estimator with the gamma kernel estimator for the marginal densities.

**Proposition 2. Uniform strong consistency of** $\hat{f}_{sp}$

*Let* $f$ *be a continuous and bounded probability density function. Under assumption A2, P1 and P2, for any compact set* $I$ *in* $[0, +\infty)$, *we have*

$$\sup_{t\in I}\left|\hat{f}_{sp}(x) - f(x)\right| \xrightarrow{a.s.} 0 \quad as \quad n \longrightarrow +\infty. \tag{14}$$

■

If we also assume a twice differentiable density function then the rate of convergence of $\hat{f}_{sp}$ can be deduced from Proposition 2. The last proposition deals with the asymptotic normality of the semiparametric density estimator. The result is useful for goodness of fit tests and confidence intervals.

**Proposition 3. Asymptotic normality of** $\hat{f}_{sp}$

*Suppose that* $f_1, ..., f_d$ *are twice differentiable at* $x$. *We suppose that the bandwidth parameters satisfy (12). Under assumption P1 and P2. we have*

$$n^{\frac{1}{2}}\left(\sum_{j=1}^{d}V_j^*(x)b_j^{-1/2}\right)^{-\frac{1}{2}}\left(\hat{f}_{sp}(x) - f(x) - \mu_x\right) \xrightarrow{\mathcal{D}} N(0,1) \tag{15}$$

*where*

$$V_j^*(x) = \begin{cases} (2\sqrt{\pi})^{-1}\gamma_\theta^2(x)\tilde{f}_j^2(x)f_j(x_j)x_j^{-1/2} & if\, x_j/b_j \to \infty \\ \\ \frac{\Gamma(2\kappa+1)}{2^{2\kappa+1}\Gamma^2(\kappa+1)}\gamma_\theta^2(x)\tilde{f}_j^2(x)f_j(x_j)b_j^{-1/2} & if\, x_j/b_j \to \kappa \end{cases} \tag{16}$$

*and*

$$\mu_x = \sum_{j=1}^{d} b_j B_j^*(x). \tag{17}$$

■

The next remark deals with the asymptotic normality of the semiparametric estimator with the local linear and the second gamma kernel estimator for the marginal densities.

**Remark 2.** *The asymptotic normality in (15) remains valid*

- *For the local linear kernel, with* $b_j = h_j^2$

$$B_j^*(x) = \gamma_\theta(x)\tilde{f}_j(x)\frac{s_2^2(p_j) - s_1(p_j)s_3(p_j)}{s_2(p_j)s_0(p_j) - s_1^2(p_j)}\frac{f''(x_j)}{2}$$

*and*

$$V_j^*(x) = \gamma_\theta(x)^2(x)\tilde{f}_j^2(x)f(x)\frac{s_2^2(p_j) - 2s_2(p_j)s_1(p_j)e_1(p_j) + s_1^2(p_j)e_2(p_j)}{(s_2(p_j)s_0(p_j) - s_1^2(p_j))^2}$$

*where* $p_j = x_j/h_j$, $s_i(p) = \int_{-1}^{p} u^i K(u)du$ *and* $e_i = \int_{-1}^{p} u^i K^2(u)du$

- *For the modified gamma kernel, with the same* $V_j^*$ *as for gamma kernel but with*

$$B_j^*(x) = \gamma_\theta(x)\tilde{f}_j(x)\begin{cases} \frac{1}{2}x_j f''(x_j) & \text{if } x_j \geq 2b_j \\ \\ \xi_{b_j}(x_j)f'(x_j) & \text{if } x_j < 2b_j \end{cases}$$

*where* $\xi_b(x) = (1-x)(\rho(2,x) - x/b)/(1 + b\rho(2,x) - x)$.

The two terms $\mu$ and $V_J^*$ are unknown since they depend on the unknown density function. In practice, we can replace the density function in these terms by the semiparametric estimator, thanks to the uniform strong convergence in Proposition (2). Remark that the presence of the term $\mu$ in (15) is due to the bias. This term disappears if we choose the bandwidth parameter $b_j = o(n^{-2/5})$ for the gamma kernels and $h_j = o(n^{-1/5})$ for the local linear kernels . Remark also that for the gamma kernels the variances increase at points near zero but decrease for points further away from zero.

## 4 Finite sample properties

For bivariate random variables, we compare the mean and the variance of the mean integrated squared error (MISE) of the semiparametric estimator via copula using the Gaussian, local linear and modified gamma kernel. The gamma kernel estimator is not considered as it performs less well than the modified kernel as documented for example in Chen (2000). We consider the Gaussian copula and the Gumbel-Hougaard copula, denoted respectively $C1$ and $C2$, which are defined as follows

$$C1(u_1,u_2)=\frac{1}{\sqrt{1-\alpha^2}}\exp\left\{\frac{-(w_1^2-2\alpha w_1w_2+w_2^2)}{2(1-\alpha^2)}\right\}\exp\left\{\frac{w_1^2+w_2^2}{2}\right\} \tag{18}$$

and

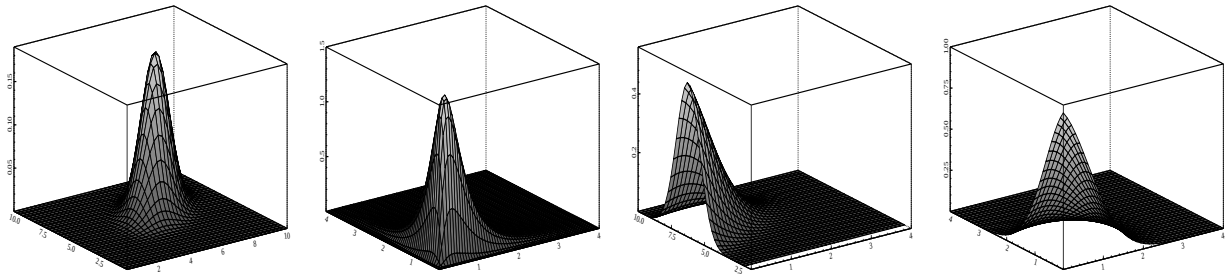
$$C2(u_1,u_2)=\frac{\exp\left(-(v_1+v_2)^{1/\beta}\right)}{u_1u_2}\frac{\{\ln(u_1)\ln(u_2)\}^{\alpha-1}\left(\{v_1+v_2\}^{1/\beta}+\beta-1\right)}{\left(-\{v_1+v_2\}^{2-1/\beta}\right)} \tag{19}$$

where $\alpha$ is the correlation coefficient, $\tau=1-\beta^{-1}$ is Kendall's tau, $w_i=\Phi^{-1}(u_i)$, $v_i=(-\ln(u_i))^{\beta}$ and $\Phi^{-1}$ is the inverse of normal distribution function. We consider four following data generating processes (the densities are displayed in Figure 3):

- Model A: no boundary problem: normal density with mean $(\mu_1,\mu_2)=(6,6)$ and variance $(\sigma_1^2,\sigma_2^2)=(1,1)$ and correlation $r=0.5$.
- Model B: independent inverse Gaussian with mean $\mu=0.8$ and the scaling parameter $\lambda=1$.
- Model C: one boundary problem: Truncated normal density with mean $(\mu_1,\mu_2)=(-0.5,6)$ and variance $(\sigma_1^2,\sigma_2^2)=(1,1)$ and correlation $r=0.5$.
- Model D: two boundary problems: Truncated normal density with mean $(\mu_1,\mu_2)=(-0.5,-0.5)$ and variance $(\sigma_1^2,\sigma_2^2)=(1,1)$ and correlation $r=0.8$.

We consider the sample sizes 250, 500 and 1000 and we perform 100 replications for each model. In each replication the bandwidth is chosen such that the integrated squared error is minimized. This theoretical bandwidth is compared with the bandwidth selected using the univariate least squares cross-validation method.

For model A and B, we report the mean and the standard deviation of the MISE in Table 1. A basic point is that the mean and the variance of the MISE are both negatively related to the



(a) No boundary problem (b) Inverse Gaussian (c) One boundary problem (d) Two boundary problems

Figure 3: Density functions considered for the simulations.

sample size. This is indeed true for all models. In fact, for model A and B, the mean MISE is lower and decreases faster with the sample size for the Gumbel-Hougaard copula. For example, for the local linear estimator and model A the mean MISE for $n = 250$ is equal to 0.0033 and 0.0032 for the Gaussian and Gumbel-Hougaard copula respectively. For $n = 1000$ this decreases respectively to 0.002 and 0.0015. However, the standard deviation for the Gumbel-Hougaard copula is slightly larger and decreases at the same rate as the Gaussian copula. The overall performance of all the estimators is similar for model A

We also find in Table 1 that given the copula for model B, here the density has more mass closer to zero, the mean and variance of MISE of the modified gamma estimator are smaller than those of the other estimators. Also, the local linear estimator performs better that the Gaussian kernel in term of mean and variance of ISE. Therefore, with respect to model B, we prefer as estimator the modified gamma with the Gumbel-Hougaard copula.

For model A and B, the theoretical bandwidths and the univariate least squares cross-validation (LSCV) implied bandwidths are reported in Table 2. As for the MISE, the bandwidths are negatively related to the data sample size, and this is uniformly true for all the models. We give first some remarks for model A. The estimator with the Gumbel-Hougaard copula uses slightly large bandwidths than those with the Gaussian copula. In terms of variance, the gumbel copula leads to a less variable bandwidth. This remark holds for the estimator with Gaussian, local linear and modified gamma kernels. The mean of the theoretical and LSCV bandwidths of the estimator with the Gaussian kernel are similar. But, the variance of the LSCV bandwidths is greater than the theoretical bandwidth. The LSCV rule selects bandwidths which are in general smaller than the theoretical bandwidth for the local linear kernel and larger for modified gamma kernel. The LSCV

Table 1: Mean and standard deviation of $L_2$ error for the density function estimators.

| | | | Gaussian | | Local Linear | | Modified Gamma | |
|---|---|---|---|---|---|---|---|---|
| | | | C1 | C2 | C1 | C2 | C1 | C2 |
| | n=250 | Mean | 0.0034604 | 0.0032558 | 0.0033483 | 0.0031577 | 0.0034706 | 0.0033808 |
| | | Std dev | 0.0008184 | 0.0012831 | 0.0008045 | 0.0012706 | 0.0008197 | 0.0013154 |
| A | n=500 | Mean | 0.0025740 | 0.0020527 | 0.0025007 | 0.0019930 | 0.0025843 | 0.0021315 |
| | | Std dev | 0.0003590 | 0.0005381 | 0.0003554 | 0.0005306 | 0.0003607 | 0.0005464 |
| | n=1000 | Mean | 0.0021078 | 0.0014917 | 0.0020616 | 0.0014566 | 0.0021150 | 0.0015405 |
| | | Std dev | 0.0002167 | 0.0003623 | 0.0002138 | 0.0003555 | 0.0002190 | 0.0003688 |
| | n=250 | Mean | 0.0175830 | 0.0163880 | 0.0162330 | 0.0149720 | 0.0154400 | 0.0121850 |
| | | Std dev | 0.0056233 | 0.0059273 | 0.0052316 | 0.0054240 | 0.0057964 | 0.0055539 |
| B | n=500 | Mean | 0.0127010 | 0.0102840 | 0.0116060 | 0.0093563 | 0.0104410 | 0.0070886 |
| | | Std dev | 0.0041244 | 0.0045425 | 0.0036716 | 0.0041353 | 0.0034800 | 0.0032665 |
| | n=1000 | Mean | 0.0092182 | 0.0057773 | 0.0083068 | 0.0052725 | 0.0080871 | 0.0040154 |
| | | Std dev | 0.0023277 | 0.0025010 | 0.0020823 | 0.0022283 | 0.0018718 | 0.0017529 |

A: bivariate normal, B: two independent inverse Gaussian. Std dev: standard deviation. Copula1: Gaussian copula and Copula2: Gumbel-hougaard copula

rule seems to performs better for the modified gamma kernel than the local linear kernel. Also, with respect to the variance, the LSCV bandwidths are less stable than the theoretical bandwidth. For example for the modified gamma kernel with $n = 500$ the mean theoretical bandwidths for the Gumbel-Hougaard copula are $(0.014, 0.013)$ and the mean LSCV bandwidths are $(0.015, 0.016)$. The standard deviation for the those bandwidths are respectively $(0.0054, 0.0048)$ and $(0.0068, 0.0072)$.

For model B, the theoretical bandwidths are the almost the same for Gaussian kernel with the two considered copulas. The means of LSCV bandwidths are slightly larger than the theoretical bandwidth. In term of variance they are similar. From model A to B, the estimator with the Gaussian kernel and local linear kernel uses small bandwidths, whereas the modified gamma kernel uses slightly large bandwidths. It seems that the two first estimator try to reduce the bias and the last one try to reduce the variance. We also remark in general for both models A and B, the behavior of the first and second bandwidth are similar since the densities under study are quite symmetric. This changes in the case of one boundary problem in model C.

For model C and D, with pronounced boundary problems, we report the mean and the variance of the MISE in Table 3. We do not consider the Gaussian kernel as it suffers from the boundary bias. Given the copula, the estimator with modified gamma kernel dominates slightly in terms of mean MISE. Also, the modified gamma kernel performs better in term of variance. The Gumbel-Hougaard copula seems to be more adequate than the Gaussian copula for both the local linear and modified gamma kernels. For example, for $n = 250$, the mean integrated of the estimator with modified gamma kernel is 0.010251 and for Gumbel-Hougaard copula it is 0.0057428. From model C to D, that is when the concentration of observations becomes large in the boundary region, the mean and the variance of the MISE increase.

For model C and D, the theoretical bandwidths and the univariate least squares cross-validation (LSCV) implied bandwidths are reported in Table 4. The estimator with gumbel-Hougaard copula uses larger bandwidths than the Gaussian copula and LSCV for Local linear kernel. The univariate LSCV rule yields closer results with respect to the theoretical bandwidths for the estimator with the modified gamma kernel than the one with the local linear kernel. The variance of the univariate LSCV implied bandwidths is in general smaller for the estimator with the local linear kernel in both models. However, the variance is larger for the modified gamma for model D. We conclude that also for models C and D the modified gamma Gumbel-Hougaard semiparametric estimator is the best configuration.

Table 2: Mean and standard deviation of theoretical and LSCV Bandwidths for the semiparamatric estimator with Gaussian copula.

| | | | Gaussian | | local linear | | Modified Gamma | |
|---|---|---|---|---|---|---|---|---|
| | | | Mean | Std dev | Mean | Std dev | Mean | Std dev |
| | n=250 | C1 | (0.3267,0.3059) | (0.0613,0.0609) | (0.7115,0.6717) | (0.1576,0.1433) | (0.0185,0.0164) | (0.0071,0.0064) |
| | | C2 | (0.3377,0.3211) | (0.0592,0.0529) | (0.7376,0.6972) | (0.1355,0.1245) | (0.0189,0.0165) | (0.0067,0.0057) |
| | | LSCV | (0.3299,0.3537) | (0.1124,0.1069) | (0.5256,0.5571) | (0.3074,0.3438) | (0.0207,0.0240) | (0.0101,0.0098) |
| A | n=500 | C1 | (0.2822,0.2718) | (0.0545,0.0535) | (0.6014,0.5790) | (0.1462,0.1493) | (0.0134,0.0123) | (0.0057,0.0055) |
| | | C2 | (0.2925,0.2857) | (0.0551,0.0519) | (0.6431,0.6257) | (0.1232,0.1168) | (0.0142,0.0132) | (0.0054,0.0048) |
| | | LSCV | (0.2943,0.2943) | (0.0851,0.0872) | (0.4124,0.3956) | (0.2422,0.2410) | (0.0155,0.0162) | (0.0068,0.0072) |
| | n=1000 | C1 | (0.2355,0.2378) | (0.0421,0.0440) | (0.4845,0.5025) | (0.1255,0.1342) | (0.0097,0.0099) | (0.0032,0.0035) |
| | | C2 | (0.2510,0.2516) | (0.0463,0.0412) | (0.5454,0.5610) | (0.1062,0.0947) | (0.0101,0.0104) | (0.0034,0.0034) |
| | | LSCV | (0.2670,0.2504) | (0.0669,0.0749) | (0.3458,0.3315) | (0.1943,0.1874) | (0.0128,0.0123) | (0.0045,0.0045) |
| | n=250 | C1 | (0.0938,0.0960) | (0.0252,0.0272) | (0.1976,0.2055) | (0.0533,0.0555) | (0.0229,0.0249) | (0.0099,0.01120) |
| | | C2 | (0.0944,0.0984) | (0.0199,0.0235) | (0.2080,0.2144) | (0.0525,0.0588) | (0.0244,0.0260) | (0.0099,0.01180) |
| | | LSCV | (0.1013,0.1022) | (0.0249,0.0297) | (0.1841,0.1781) | (0.0726,0.0756) | (0.0283,0.0296) | ( 0.0077,0.0076) |
| B | n=500 | C1 | (0,0794,0,0823) | (0.0245,0.0184) | (0.1692,0.1795) | (0.0527,0.0414) | (0.0169,0.0161) | (0.00947,0.0087) |
| | | C2 | (0.0789,0.0830) | (0.0209,0.0178) | (0.1771,0.1804) | (0.0475,0.0412) | (0.0178,0.0176) | (0.00832,0.0083) |
| | | LSCV | (0.0901,0.0948) | (0.0228,0.0183) | (0.1637,0.1860) | (0.0619,0.0553) | (0.0245,0.0249) | ( 0.0068,0.0071) |
| | n=1000 | C1 | (0.0731,0.0697) | (0.0185,0.0197) | (0.1562,0.1529) | (0.0395,0.0433) | (0.0125,0.0131) | (0.00654,0.0066) |
| | | C2 | (0.0691,0.0685) | (0.0158,0.0163) | (0.1450,0.1478) | (0.0387,0.0360) | (0.0131,0.0134) | (0.00577,0.0064) |
| | | LSCV | (0.0852,0.0858) | (0.0187,0.0187) | (0.1613,0.1632) | (0.0413,0.0458) | (0.0212,0.0211) | (0.0034,0.0030) |

A: bivariate normal, B: two independent inverse Gaussian. Std dev: standard deviation. C1: Gaussian copula and C2: Gumbel-hougaard copula

Table 3: Mean and standard deviation of $L_2$ error for the density function estimators.

| | | | local linear | | Modified Gamma | |
|---|---|---|---|---|---|---|
| | | | C1 | C2 | copula1 | copula2 |
| | n=250 | Mean | 0.012974 | 0.0079014 | 0.010251 | 0.0057428 |
| | | Std dev | 0.003387 | 0.0028255 | 0.003392 | 0.0025296 |
| C | n=500 | Mean | 0.010664 | 0.0056661 | 0.008506 | 0.0041128 |
| | | Std dev | 0.002698 | 0.0019947 | 0.002478 | 0.0016608 |
| | n=1000 | Mean | 0.009642 | 0.0043516 | 0.007751 | 0.0031983 |
| | | Std dev | 0.001666 | 0.0011871 | 0.001596 | 0.0010523 |
| | n=250 | Mean | 0.038121 | 0.021344 | 0.031011 | 0.015433 |
| | | Std dev | 0.008557 | 0.005397 | 0.008063 | 0.005286 |
| D | n=500 | Mean | 0.032594 | 0.015909 | 0.027420 | 0.011366 |
| | | Std dev | 0.006742 | 0.003951 | 0.005949 | 0.003374 |
| | n=1000 | Mean | 0.028969 | 0.012132 | 0.025543 | 0.009474 |
| | | Std dev | 0.005695 | 0.003006 | 0.005362 | 0.002471 |

C: truncated bivariate normal with one boundary problem, D:truncated bivariate normal with two boundary problems. Std dev: standard deviation. Copula1: Gaussian copula and Copula2: Gumbel-hougaard copula

Table 4: Mean and standard deviation of theoretical and LSCV Bandwidths for the semiparamatric estimator with Gaussian copula.

| | | | local linear | | Modified Gamma | |
|---|---|---|---|---|---|---|
| | | | Mean | Std dev | Mean | Std dev |
| | n=250 | C1 | (0.1442,0.6034) | (0.0698,0.1520) | (0.0694,0.0120) | (0.0560,0.0056) |
| | | C2 | (0.1987,0.6642) | (0.1013,0.1329) | (0.0801,0.0159) | (0.0513,0.0069) |
| | | LSCV | (0.1641,0.4193) | (0.0689,0.2643) | (0.0766,0.0173) | (0.0441,0.0070) |
| C | n=500 | C1 | (0.1249,0.4609) | (0.0548,0.1046) | (0.0718,0.0076) | (0.0461,0.0029) |
| | | C2 | (0.1507,0.5468) | (0.0540,0.1275) | (0.0727,0.0110) | (0.0364,0.0041) |
| | | LSCV | (0.1477,0.3769) | (0.0615,0.2362) | (0.0689,0.0126) | (0.0299,0.0056) |
| | n=1000 | C1 | (0.1042,0.3778) | (0.0456,0.1171) | (0.0784,0.0052) | (0.0477,0.0017) |
| | | C2 | (0.1309,0.4633) | (0.0449,0.1008) | (0.0697,0.0081) | (0.0335,0.0027) |
| | | LSCV | (0.1315,0.3129) | (0.0496,0.2106) | (0.0584,0.0099) | (0.0324,0.0041) |
| | n=250 | C1 | (0.1989,0.1633) | (0.0931,0.0929) | (0.0502,0.0443) | (0.0285,0.0289) |
| | | C2 | (0.2379,0.2051) | (0.0954,0.0939) | (0.0554,0.0501) | (0.0299,0.0321) |
| | | LSCV | (0.1321,0.1237) | (0.0618,0.0570) | (0.0696,0.0769) | (0.0380,0.0379) |
| D | n=500 | C1 | (0.1346,0.1212) | (0.0573,0.0555) | (0.0409,0.0352) | (0.0287,0.0229) |
| | | C2 | (0.1610,0.1459) | (0.0630,0.0578) | (0.0415,0.0389) | (0.0238,0.0214) |
| | | LSCV | (0.1091,0.1086) | (0.0467,0.0377) | (0.0616,0.0677) | (0.0310,0.0287) |
| | n=1000 | C1 | (0.1057,0.1054) | (0.0462,0.0450) | (0.0335,0.0358) | (0.0244,0.0239) |
| | | C2 | (0.1245,0.1242) | (0.0474,0.0445) | (0.0366,0.0357) | (0.0218,0.0211) |
| | | LSCV | (0.0952,0.1001) | (0.0361,0.0327) | (0.0550,0.0552) | (0.0244,0.0233) |

C: truncated bivariate normal with one boundary problem, D:truncated bivariate normal with two boundary problems. Std dev: standard deviation. Copula1: Gaussian copula and Copula2: Gumbel-hougaard copula

Finally, we also give the mean and standard deviations for the copula parameters. Table 5 and 6 reports the correlation coefficient of the Gaussian copula and the Kendall's tau of the Gumbel-Hougaard copula. From these tables we can for example see that both the correlation and the

Table 5: Mean and standard deviation of theta parameter for Gaussian copula.

| | | T=250 | T=500 | T=1000 |
|---|---|---|---|---|
| A | Mean | 0.49866 | 0.49750 | 0.50065 |
| | Standard deviation | 0.04127 | 0.03205 | 0.02317 |
| B | Mean | 0.00697 | 0.00050 | -0.00238 |
| | Standard deviation | 0.06417 | 0.04269 | 0.03159 |
| C | Mean | 0.28569 | 0.29145 | 0.29048 |
| | Standard deviation | 0.06529 | 0.04117 | 0.02998 |
| D | Mean | 0.52692 | 0.53138 | 0.53231 |
| | Standard deviation | 0.04948 | 0.03286 | 0.02578 |

A: normal, B: Inverse Gaussian, C: truncated normal (one boundary problem), D: truncated normal (two boundary problems).

Kendall's tau are close to zero model A and that the standard deviations decrease with the sample size. Note that the correlation for model D is not underestimated since the dependence reduces because of the truncation at the origin.

# 5 Application

We collect data for 558 companies from Compustat for the year 2005. The first variable (Compustat item 24, denoted C24) is the price of the stock of the company when the books are closed at the end of the accounting year with mean 75.167, standard deviation 103.13 and skewness 2.1295. The second variable (Compustat item 25,denoted C25) is the number of shares that can be bought on the stock market with mean 21.953, standard deviation 20.302 and skewness 1.1237. The correlation between the two variables is 0.33392. Figure 4 shows the scatter plot and the semiparametric density estimates using the Gumbel-Hougaard copula with modified gamma kernels where the bandwidth parameters are selected by the univariate LSCV method and are respectively equal to $b_1 = 0.15$

Table 6: Mean and standard deviation of theta parameter for Gumbel-Hougaard copula.

| | | T=250 | T=500 | T=1000 |
|---|---|---|---|---|
| A | Mean | 0.33433 | 0.33546 | 0.33615 |
| | Standard deviation | 0.03711 | 0.02922 | 0.01681 |
| B | Mean | 0.00843 | 0.00554 | 0.00532 |
| | Standard deviation | 0.04099 | 0.03105 | 0.02042 |
| C | Mean | 0.17811 | 0.18028 | 0.17997 |
| | Standard deviation | 0.04462 | 0.02990 | 0.02057 |
| D | Mean | 0.31227 | 0.3136 | 0.31597 |
| | Standard deviation | 0.03589 | 0.0241 | 0.01949 |

A: normal, B: Inverse Gaussian, C: truncated normal (one boundary problem), D: truncated normal (two boundary problems).

and $b_2 = 4$. We also show the estimated marginal densities that constitute the semiparametric estimator. The Kendall's tau is equal to 0.2423. We remark a high concentration close to the origin, hence the Gaussian kernel is not consistent for such data.

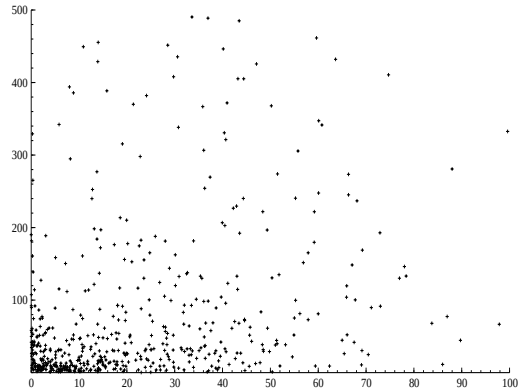
We investigate next the behavior of these two variables conditional on current assets (Compustat item 4, denoted C4) and on dividends per share by ex-date (Compustat item 26, denoted C26) by comparing the densities. Figure 5 displays the semiparametric estimator with the modified gamma kernel for densities of C24-C25 for companies with zero dividends and zero debt and the density of C24-C25 with positive dividends and positive debts. For the densities conditional to dividends it is visually clear that they are different. However, for the densities conditional to debt it is less obvious if they are different. Therefore, we perform the following test

$$H_0: \quad f(x,y|Z=0) = f(x,y|Z=1), \quad H_1: \quad f(x,y|Z=0) \neq f(x,y|Z=1) \tag{20}$$

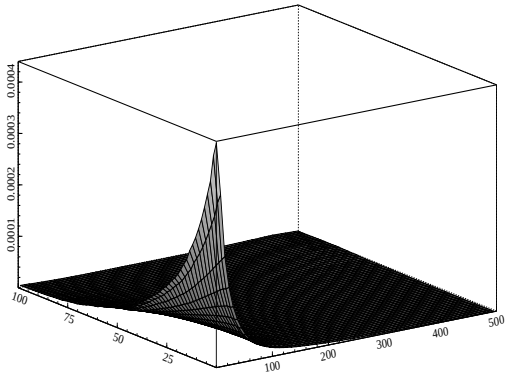
where $f(x,y|Z=0)$ (resp. $f(x,y|Z=1)$) is the joint density of C24-C25 for companies with zero debt. We consider as test statistic

$$T_1 = \sup |f(x,y|Z=0) - f(x,y|Z=1)|.$$

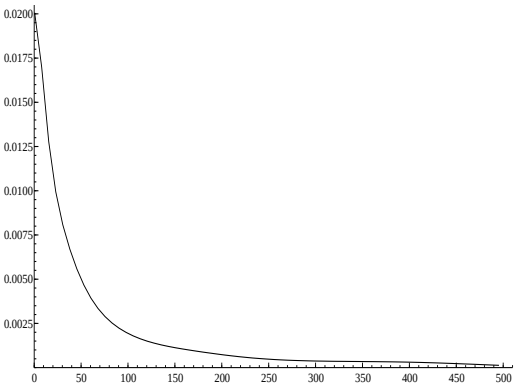
To evaluate the P-value of the test we use the nonparametric bootstrap by doing $B = 5000$ replications. We did not consider the following test statistic $T_2 = \int \left(f(x,y|Z=0) - f(x,y|Z=1)\right)^2 dxdy$



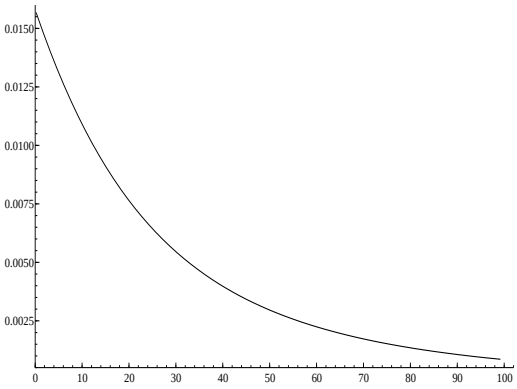

(a) Scatter plot: C24-C25




(b) Gamma kernel estimator for C24-C25 density

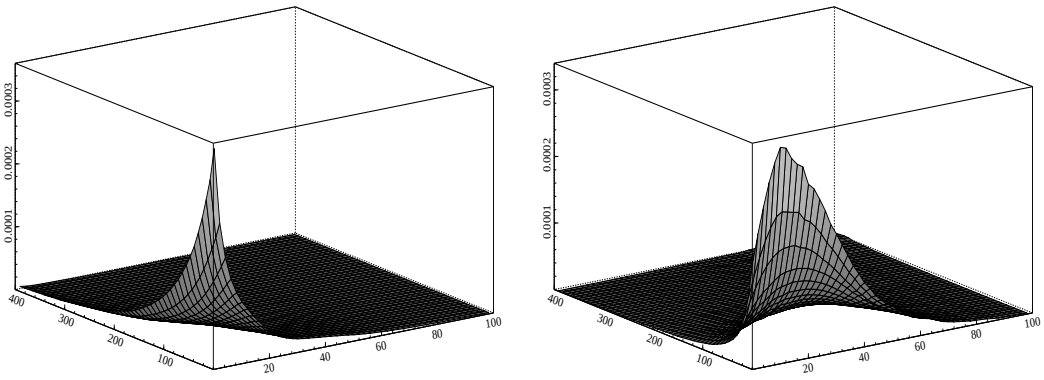



(c) Gamma kernel estimator for C24 density



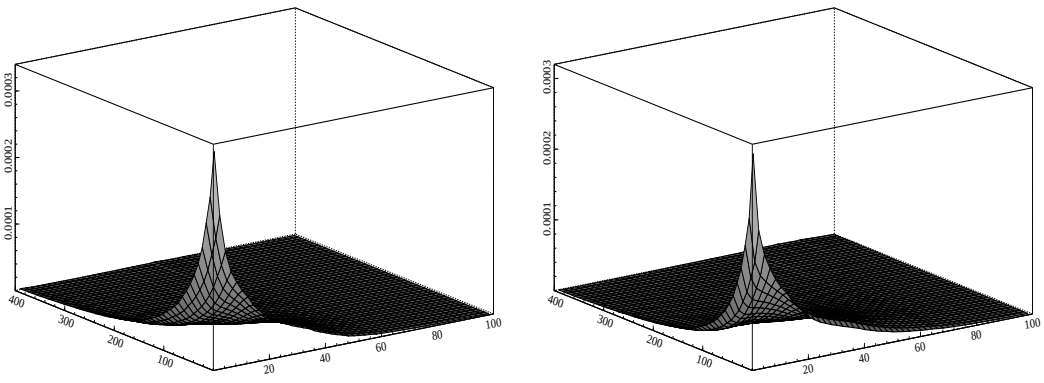

(d) Gamma kernel estimator for C25 density

Figure 4: Scatter plot and gamma kernel density estimator for C24-C25 data. The bottom: gamma kernel density estimator for the C24 and C25




(a) Gamma kernel estimator density for C24-C25 with dividend=0

(b) Gamma kernel estimator density for C24-C25 with dividend$>0$




(c) Gamma kernel estimator density for C24-C25 with debt=0

(d) Gamma kernel estimator density for C24-C25 with debt$>0$

Figure 5: Gamma kernel density estimator for C24-C25 data conditional to dividends and debt

since the nonparametric bootstrap is not consistent here, see Rémillard and Scaillet (2006). For more details on bootstrap conditions see also Bickel and Freedman (1981) and Bickel, Götze, and Zwet (1997). The p-value for the test is 0.4646 so we do not reject the null hypothesis that both densities are the same.

# 6 Conclusion

This paper proposes a multivariate semiparametric density estimation method which is robust to both the boundary and the curse of dimension problem. The estimator combines gamma or local linear kernels the support of which matches that one of the underlying multivariate density, and semiparametric copulas. This leads to an estimator which is easy to implement. We derive asymptotic properties such as the mean integrated squared error, uniform strong consistency and asymptotic normality. In the simulations, we compare the finite sample performance of the (modified) gamma and the local linear estimator for the marginal densities using the Gaussian and the Gumbel-Hougaard copula. We find that the models in the simulation study are preferably estimated using the modified gamma Gumbel-Hougaard semiparametric estimator. We also learn from the simulations that the univariate least squares cross validation technique to select bandwidths for the marginal density estimators works well. Therefore, bandwidth selection for our estimator can be done in a computational straightforward manner. In the application, we estimate the joint density of the stock price and the total number of shares outstanding using data of 558 US companies observed in 2005 and we test if the density depends on the fact that dividends are paid out or not, and on the fact that there is debt outstanding or not.

# Appendix

We give the proofs for the semiparametric estimator using the gamma kernel estimator.

**Proof of proposition 1**

The proof of proposition 1 is straightforward from the proof of the result on the mean squared error in Liebscher (2005) for the standard kernel and with the same bandwidth, the bias and the variance of the gamma kernel in the univariate case and the fact that

$$K_G(b,t)(x) \leq \sqrt{\frac{1}{2\pi x b}}. \tag{21}$$

■

**Proof of Proposition 2**

The semiparametric estimator can be expressed as:

$$\hat{f}_{sp}(x) = f(x) + \gamma_\theta(F_1(x_1), ..., F_d(x_d)) \sum_{j=1}^{d} \Big( \hat{f}_j(x_j) - f_j(x_j) \Big) \prod_{l=1}^{j-1} f_l(x_l) \prod_{k=j+1}^{d} \hat{f}_k(x_k) + \bar{\gamma}(x) \prod_{j=1}^{d} \hat{f}_j(x_j) \quad (22)$$

where $\prod_{j=1}^{0} = \prod_{j=d+1}^{d} = 1$ and

$$\bar{\gamma}(x) = \gamma_{\hat{\theta}}(F_{n1}(x_1), ..., F_{nd}(x_d)) - \gamma_\theta(F_1(x_1), ..., F_d(x_d)).$$

Under the continuity of the distribution functions $F_1(x_1), ..., F_d(x_d)$

$$\sup_{x_j} \Big( |\hat{F}_{nj}(x_j) - F_j(x_j)| \Big) \xrightarrow{a.s.} 0. \quad \text{for} j = 1, ..., d. \quad (23)$$

Under assumption P1 and P2, and 23 we have,

$$\sup_{x \in I} \left( |\bar{\gamma}(x)| \right) \xrightarrow{a.s.} 0.$$

Hence, using the uniform weak consistence of $\hat{f}_j (j = 1, ..., d)$

$$\sup_{x} \left( \bar{\gamma}(x) \prod_{j=1}^{d} \hat{f}_j(x_j) \right) \xrightarrow{a.s.} 0. \quad (24)$$

From Bouezmarni and Scaillet (2005), under assumption B3 and the continuity of density functions $f_i, ..., f_d$ we have

$$\sup_{x_j} \Big( |\hat{f}_j(x_j) - f_j(x_j)| \Big) \xrightarrow{a.s.} 0. \quad \text{for} \quad i = 1, ..., d.$$

Therefore

$$\sup_{x} \left( \gamma_\theta \sum_{j=1}^{d} \Big( \hat{f}_j(x_j) - f_j(x_j) \Big) \prod_{l=1}^{j-1} f_l(x_l) \prod_{k=j+1}^{d} \hat{f}_k(x_k) \right) \xrightarrow{a.s.} 0. \quad (25)$$

The uniform strong consistency of the semiparametric estimator with gamma kernel can be deduced from (22), (24) and (25).

■

**Proof of Proposition 3**

From assumption P1 and P3 and using the consistency of the empirical distribution $\hat{F}_{nj}$ and the density estimators $\hat{f}_j$, $j = 1, ..., d$, we have

$$\left| \bar{\gamma}(x) \prod_{j=1}^{d} \hat{f}_j(x_j) \right| = O_P(n^{-1/2}\sqrt{\ln(n)}). \tag{26}$$

Therefore and from (22),

$$\begin{aligned} |\hat{f}_{sp}(x) - f(x)| &= \gamma_\theta(F_1(x_1), ..., F_d(x_d)) \sum_{j=1}^{d} \left( \hat{f}_j(x_j) - f_j(x_j) \right) \prod_{l=1}^{j-1} f_l(x_l) \prod_{k=j+1}^{d} \hat{f}_k(x_k) \\ &\quad + O_P(n^{-1/2}\sqrt{\ln(n)}) \\ &= \gamma_\theta(F_1(x_1), ..., F_d(x_d)) \left\{ \sum_{i=1}^{n} U_i + \sum_{j=1}^{d} \left( E(\hat{f}_j(x_j)) - f_j(x_j) \right) \tilde{f}_j(x) \right\} \Psi_j \quad (27) \\ &\quad + O_P(n^{-1/2}\sqrt{\ln(n)}) \end{aligned}$$

where

$$U_i = \frac{1}{n} \sum_{j=1}^{d} \left( K_G(b_j, X_i^j)(x_j) - E\left( K_G(b_j, X_i^j)(x_j) \right) \right) \tilde{f}_j(x)$$

and

$$\Psi_j = \prod_{l=j+1}^{d} \hat{f}_l(x_l) f_l(x_l)^{-1}.$$

Note that using the consistency of $\hat{f}_l(x_l)$ we get

$$\Psi_j \xrightarrow{P} 1, \quad \text{for } \ j = 1, ..., d. \tag{28}$$

Denote $\alpha = n^{1/2} \left( \sum_{j=1}^{d} V_j^* b_j^{-1/2} \right)^{-1/2}$.

Using the expectation of the gamma kernel estimator in the univariate case

$$\sum_{j=1}^{d} \left( E(\hat{f}_j(x_j)) - f_j(x_j) \right) \tilde{f}_j(x) = \sum_{j=1}^{d} b_j (f_j'(x_j) + \frac{1}{2} x_j f_j''(x_j)) \tilde{f}_j(x) + O(\sum_{j=1}^{d} b_j^2). \tag{29}$$

Hence, from (12) and by omitting $F_j$ in $\gamma_\theta(F_1(x_1), ..., F_d(x_d))$

$$\alpha \left( \gamma_\theta \sum_{j=1}^{d} \left( E(\hat{f}_j(x_j)) - f_j(x_j) \right) \tilde{f}_j(x) \right) = \alpha \sum_{j=1}^{d} b_j B_j + O(n^{-\frac{2}{5}}). \tag{30}$$

Now, it remains to prove

$$S_n = \alpha\gamma_\theta \sum_{i=1}^{n} U_i \xrightarrow{\mathcal{D}} N(0,1). \tag{31}$$

To do this, we apply Liapunov central limit proposition to independent random variables $V_i = \alpha\gamma_\theta U_i$ and show that $\mathrm{Var}(S_n) = 1 + o(1)$ and $\lim_n \sum_{i=1}^{n} \boldsymbol{E}|V_i|^3 = 0$. We calculate the variance of $U_i$.

$$\begin{aligned}
\mathrm{Var}(U_i) &= \frac{1}{n^2}\sum_{j=1}^{d} \tilde{f}_j^2(x)\mathrm{Var}\left(K_G(b_j, X_i^j)(x_j)\right) \\
&+ \frac{2}{n^2}\sum_{j=1}^{d}\sum_{l>j} \tilde{f}_j(x)\tilde{f}_l(x)\mathrm{Cov}\left(K_G(b_j, X_i^j)(x_j), K_G(b_l, X_i^l)(x_l)\right)
\end{aligned}$$

On the one hand,

$$\mathrm{Var}\left(K_G(b_j, X_i^j)(x_j)\right) = (2\sqrt{\pi})^{-1} b_j^{-1/2} f_j(x_j) x_j^{-1/2} \tag{32}$$

On the other hand, we can show that

$$\mathrm{Cov}\left(K_G(b_j, X_i^j)(x_j), K_G(b_l, X_i^l)(x_l)\right) = O(1). \tag{33}$$

Therefore,

$$\mathrm{Var}(S_n) = \frac{\alpha^2\gamma_\theta^2}{n}\sum_{j=1}^{d} \tilde{f}_j^2(x)(2\sqrt{\pi})^{-1} b_j^{-1/2} f_j(x_j) x_j^{-1/2} + o(1). \tag{34}$$

Now, using inequality (21), the variance of $K_G(b_j, X_i^j)(x_j)$ in (32) and (12)

$$\begin{aligned}
E|V_i|^3 &\leq \frac{\alpha^3\gamma_\theta^3}{n^3}\sum_{j=1}^{d} \tilde{f}_j^3(x) \int K_G^3(b_j, t)(x_j) f_j(t) dt \\
&= O(n^{-\frac{7}{5}}).
\end{aligned} \tag{35}$$

Hence,

$$\lim_n \sum_{i=1}^{n} E|V_i|^3 = O(n^{-2/5}) \tag{36}$$

Therefore we have the asymptotic normality of $S_n$. Proposition 3 can be deduced from (27), (28), (30) and(31).

■

## References


Bickel, P. J., and D. Freedman (1981): "Some asymptotic theory for the boostrap," *Annals of Statistics*, 9, 1196–1217.

Bickel, P. J., F. Götze, and W. R. Zwet (1997): "Resampling fewer than n observations: gains, losses, and remedeies for losses," *Statistica sinica*, 7, 1–31.

Bouezmarni, T., and J. Rombouts (2006): "Nonparametric Density Estimation for Positive Time Serie," *CORE discusion paper 2006/85*.

Bouezmarni, T., and O. Scaillet (2005): "Consistency of Asymmetric Kernel Density Estimators and Smoothed Histograms with Application to Income Data," *Econometric Theory*, 21, 390–412.

Chen, S. (2000): "Probability Density Functions Estimation Using Gamma Kernels," *Annals of the Institute of Statistical Mathematics*, 52, 471–480.

Chen, S. X., and T. Huang (2007): "Nonparametric Estimation of Copula Functions for Dependent Modeling," *Canadian Journal of Statistics*, 35, 265–282.

Cho, M. (1998): "Ownership Structure, Investment, and the Corporate Value: An Empirical Analysis," *Journal of Financial Economics*, 47, 103–121.

Deheuvels, P. (1979): "La fonction de dépendance enpirique et ses propriétés. Un test non paramtétrique díndependance," *Bullettin de l'académie Royal de Belgique, Classe des Sciences*, pp. 274–292.

Genest, C., K. Ghoudi, and L. Rivest (1995): "A semiparametric estimation procedure od dependance parameters in multivariate families of distributions," *Biometrika*, 82, 543–552.

Genest, C., and L. Rivest (1993): "Statistical inference procedure for bivariate Archimidean copulas," *Journal of the American Statistical Association*, 88, 1034–1043.

Gijbels, I., and J. Mielniczuk (1990): "Estimating The Density of a Copula Function," *Communications in Statistics - Theory and Methods*, 19, 445–464.

Grullon, G., and R. Michaely (2002): "Dividens, Share Repurchases, and the Substitution Hypothesis," *The Journal of Finance*, 57, 1649–1684.

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

Jones, M. (1993): "Simple Boundary Correction for Kernel Density Estimation," *Statistical Computing*, 3, 135–146.

Jones, M., and P. Foster (1996): "A Simple Nonnegative Boundary Correction Method for Kernel Density Estimation," *Statistica Sinica*, 6, 1005–1013.

Kim, G., M. Silvapulle, and P. Silvapulle (2007): "comparison of the semiparametric and parametric methods for estimating copulas.," *Computational Statistics and Data Analysis*, 51, 2836–2850.

Lejeune, M., and P. Sarda (1992): "Smooth Estimators of Distribution and Density Functions," *Computational Statistics and Data Analysis*, 14, 457–471.

Liebscher, E. (2005): "Semiparametric Density Estimators Using Copula," *Communications in Statistics - Theory and Methods*, 67, 318–348.

Marron, J., and D. Ruppert (1994): "Transformations to reduce boundary bias in kernel density estimation," *Journal of the Royal Statistical Society, Series B*, 56, 653–671.

Müller, H. (1991): "Smooth Optimum Kernel Estimators near Endpoints," *Biometrika*, 78, 521–530.

Müller, H., and U. Stadtmüller (1999): "Multivariate Boundary Kernels and a Continuous Least Squares Principle," *Journal of the Royal Statistical Society, Series B*, 61, 439–458.

Nelson, R. (1999): *An Introduction to Copulas. Lecture Notes in Statistics*, vol. 139. Springer, New York.

Oakes, D. (1982): "A model for association in bivariate survival data," *Journal of the Royal Statistical Society, Series B*, 44, 414–422.

——— (1986): "Semiparametric inference in a model for association in bivaraite survival data," *Biometrika*, 73, 353–361.

Rémillard, B., and O. Scaillet (2006): "Testing for Equality Between Two Copulas," *Cahiers du GERAD, G-2006-31.*

Rice, J. (1984): "Boundary Modification for Kernel Regression," *Communications in Statistics - Theory and Methods*, 13, 893–900.

Rödel, E. (1987): "R-estimation of normed bivariate density functions," *Statistics*, 18, 573–585.

Romano, C. (2002): "Calibrating and simulating copula functions :an application to the Italian stock market," *Working Paper of CIDEM, universitá degli Studi di Roma*, 12.

Sancetta, A., and S. Satchell (2004): "The Bernstein copula and its applications to modeling and approximating of multivariate distributions," *Econometric Theory*, 20, 535–562.

Schuster, E. (1985): "Incorporating Support Constraints into Nonparametric Estimators of Densities," *Communications in Statistics - Theory and Methods*, 14, 1123–1136.

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