ESTIMATING A BIVARIATE DENSITY WHEN THERE ARE EXTRA DATA ON ONE OR BOTH COMPONENTS

Peter Hall[1] Natalie Neumeyer[1,2,3]

ABSTRACT. Assume we have a dataset, $\mathcal{Z}$ say, from the joint distribution of random variables X and Y, and two further, independent datasets, $\mathcal{X}$ and $\mathcal{Y}$, from the marginal distributions of X and Y, respectively. We wish to combine $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, so as to construct an estimator of the joint density. This problem is readily solved in some parametric circumstances. For example, if the joint distribution were normal then we would combine data from $\mathcal{X}$ and $\mathcal{Z}$ to estimate the mean and variance of X; proceed analogously to estimate the mean and variance of Y; but use data from $\mathcal{Z}$ alone to estimate $E(XY)$. However, the problem is more difficult in a nonparametric setting. There we suggest a copula-based solution, which has potential benefits even when the marginal datasets $\mathcal{X}$ and $\mathcal{Y}$ are empty. For example, if the copula density is sufficiently smooth in the region where we wish to estimate it, then the effective dimension of the structure that links the marginal distributions is relatively low, and the joint density of X and Y can be estimated with a high degree of accuracy. Similar improvements in performance are available if the marginals are close to being independent. We suggest using wavelet estimators to approximate the copula density, which in cases of statistical interest can be unbounded along boundaries. Our techniques are also useful for solving recently-considered related problems, for example where the marginal distributions are determined by parametric models. Therefore the methodology has application beyond the context which motivated it. The methodology is also readily extended to more general multivariate settings.

KEYWORDS. Copula, Dimension reduction, Independence, Kernel method, Prediction, Threshold, Wavelet.

SHORT TITLE. Estimating a bivariate distribution.

---

[1] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

[2] Ruhr-Universitaet Bochum, Fakultaet fuer Mathematik, 44780 Bochum, Germany

[3] The financial support of the Deutsche Forschungsgemeinschaft is gratefully acknowledged (Sonderforschungsbereich 475 and Research Grant).

1. INTRODUCTION

Assume we observe data $\mathcal{Z} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ from the joint distribution of the bivariate random variable $Z = (X, Y)$, and that we have additional data $\mathcal{X} = \{X_{n+1}, \ldots, X_{n+p}\}$ on X, and $\mathcal{Y} = \{Y_{n+1}, \ldots, Y_{n+q}\}$ on Y. The samples $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are totally independent, as too are the data within them. We wish to estimate aspects of the joint distribution of X and Y, for example the joint density f_{XY}, or the density $f_{Y|X}$ of Y given X.

A different but related problem was treated by Spiegelman and Park (2003). These authors considered the case where only $\mathcal{Z}$ was available, but the marginal distributions of X and Y were known up to parametric models, although estimation of the joint distribution required a nonparametric approach. Spiegelman and Park's technique was to use maximum-likelihood estimation in the marginal models to estimate quantiles of the marginal distributions, and then compute a multivariate estimator of the joint distribution by using the estimated quantiles and their concomitants, rather than the original multivariate data.

Further related work includes that of Schuster and Yakowitz (1985), Olkin and Spiegelman (1987) and Jones (1993). The copula-based approach that we shall suggest, for handling the different problem considered in this paper, can be used to give alternative solutions to the problems considered there.

The copula method permits convergence-rate improvements, relative to standard rates for nonparametric bivariate inference, even if the samples $\mathcal{X}$ and $\mathcal{Y}$ are empty. In particular, if the marginals of (X, Y) are approximately independent then the copula density is close to the constant 1, and can be estimated particularly accurately. As a result, the bivariate problem of estimating f_{XY} can have a solution that admits univariate convergence rates. This also holds true if the difference between the copula density and its counterpart under the assumption of independence is sufficiently smooth; approximate independence is not necessary.

In both cases the improved convergence rates are attainable using empirical, adaptive methods. Information about the strength of dependence, or smoothness of the copula, is evaluated from the data using a threshold-based approach. More generally, the matter of empirical smoothing-parameter choice is relatively straightforward when using the copula method. Estimators of the marginal distributions can be constructed using standard techniques, employing conventional smoothing-parameter selectors.

A related semiparametric estimator was proposed by Liebscher (2005), who combined parametric estimators for the copula with nonparametric estimators for the marginal distributions.

Nonparametric estimators of copula densities have been suggested by Gijbels and Mielniczuk (1990) and Fermanian and Scaillet (2002), who used kernel methods, and Sancetta (2003) and Sancetta and Satchell (2004), who employed techniques based on Bernstein polynomials. The wavelet methods that we suggest are quite different, partly because our aim is different from those of other authors. In statistically important cases a copula density can be unbounded at boundaries, and in such instances, wavelet methods perform relatively well, whereas conventional kernel or orthogonal-series techniques can suffer from edge effects which are quite difficult to remove.

Our methods and results extend straightforwardly to multivariate cases, where one might have data on a k-vector Z and have additional samples from the distributions of its individual components. However, notation in that setting is cumbersome, and that awkwardness obscures the simplicity of our approach.

2. METHODOLOGY

Define $\mathcal{X}' = \{X_1, \ldots, X_{n+p}\}$ and $\mathcal{Y}' = \{Y_1, \ldots, Y_{n+q}\}$, and let $\mathcal{W} = (\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ denote the pooled dataset. We shall consider three specific problems: estimation of (a) f_{XY}, (b) $f_{Y|X}$ and (c) μ_X, where $\mu_X(x) = E(Y \mid X = x)$.

Our approach is founded on a representation of a bivariate distribution in terms of its marginals:

$$F_{XY}(x, y) = \Gamma\{F_X(x), F_Y(y)\}\,,$$

where F_X, F_Y and F_{XY} denote the distribution functions of X, Y and (X, Y), respectively, and Γ, the copula, is simply the joint distribution function of $W = (U, V) = (F_X(X), F_Y(Y))$. In this notation,

$$f_{XY}(x, y) = f_X(x)\, f_Y(y)\, \gamma\{F_X(x), F_Y(y)\}\,, \quad \mu_X(x) = \int F_Y^{-1}(v)\, \gamma\{F_X(x), v\}\, dv\,,$$

where $\gamma(u, v) = (\partial^2/\partial u\, \partial v)\, \Gamma(u, v)$ is the copula density, i.e. the density of W. Taking $\hat{f}_X$, $\hat{f}_Y$, $\widehat{F}_X$ and $\widehat{F}_Y$ to be conventional estimators of f_X, f_Y, F_X and F_Y, based on the respective datasets $\mathcal{X}'$ and $\mathcal{Y}'$; and $\widehat{\gamma}$ to be an estimator of γ, computed from $\mathcal{W}$; we let

$$\hat{f}_{XY}(x,y) = \hat{f}_X(x)\,\hat{f}_Y(y)\,\widehat{\gamma}\{\widehat{F}_X(x),\widehat{F}_Y(y)\}\,, \quad \hat{f}_{Y|X} = \hat{f}_{XY}/\hat{f}_X\,, \tag{2.1}$$

$$\begin{aligned}\widehat{\mu}_X(x) &= \int \widehat{F}_Y^{-1}(v)\,\widehat{\gamma}\{\widehat{F}_X(x),v\}\,dv = \frac{1}{n+q}\sum_{i=1}^{n+q} Y_i\,\widehat{\gamma}\{\widehat{F}_X(x),\widehat{F}_Y(Y_i)\} \\ &= \frac{1}{n+q}\sum_{i=1}^{n+q} Y_{(i)}\,\widehat{\gamma}\{\widehat{F}_X(x), i/(n+q)\}\end{aligned} \tag{2.2}$$

be our estimators of f_{XY}, $f_{Y|X}$ and μ_X. In (2.2), $Y_{(i)}$ denotes the ith largest value in the dataset $\mathcal{Y}'$.

The estimators $\hat{f}_X$, $\hat{f}_Y$, $\widehat{F}_X$ and $\widehat{F}_Y$ can be standard; for example,

$$\widehat{F}_X(x) = \frac{1}{n+p}\sum_{i=1}^{n+p} I(X_i \le x)\,, \quad \hat{f}_X(x) = \frac{1}{(n+p)\,h_X}\sum_{i=1}^{n+p} K\left(\frac{x-X_i}{h_X}\right),$$

where $I(X_i \le x) = 1$ if $X_i \le x$ and equals zero otherwise, and $h_X > 0$ is a bandwidth. Alternatively, $\widehat{F}_X$ could be taken to be the distribution function corresponding to $\hat{f}_X$, although that approach can introduce unwanted biases, depending on bandwidth choice for the marginal density estimators.

The copula density, γ, is supported on only the unit square, $\mathcal{S}_0$ say, and may have jump discontinuities along the boundary. In particular, this is the case if X and Y are independent, and also in a range of other instances. The setting where (X,Y) has a bivariate normal distribution, but X and Y are not independent, illustrates relatively extreme behaviour. There, γ is unbounded along sections of the boundary of $\mathcal{S}_0$. Standard kernel and orthogonal-series techniques have difficulty coping with either discontinuities or places where the target density is unbounded. Wavelet methods, however, suffer less from aberrations in such cases.

A wavelet expansion of γ is given by

$$\gamma(u,v) = \sum_k \sum_\ell a_{j_0\ell k}\,\Phi_{j_0\ell k}(u,v) + \sum_{i=1}^{3}\sum_{j=j_0}^{\infty}\sum_k\sum_\ell b_{ij\ell k}\,\Psi_{ij\ell k}(u,v)\,,$$

where $a_{j_0\ell k} = E\{\Phi_{j_0\ell k}(U,V)\}$, $b_{ij\ell k} = E\{\Psi_{ij\ell k}(U,V)\}$,

$$\begin{aligned}\Phi_{j_0\ell k}(u,v) &= 2^{j_0}\,\rho\,\phi\big(2^{j_0}\rho u-\ell\big)\,\phi\big(2^{j_0}\rho v-k\big)\,,\\ \Psi_{1j\ell k}(u,v) &= 2^{j}\,\rho\,\phi\big(2^{j}\rho u-\ell\big)\,\psi\big(2^{j}\rho v-k\big)\,,\\ \Psi_{2j\ell k}(u,v) &= 2^{j}\,\rho\,\psi\big(2^{j}\rho u-\ell\big)\,\phi\big(2^{j}\rho v-k\big)\,,\\ \Psi_{3j\ell k}(u,v) &= 2^{j}\,\rho\,\psi\big(2^{j}\rho u-\ell\big)\,\psi\big(2^{j}\rho v-k\big)\,,\end{aligned}$$

ϕ and ψ are "father" and "mother" wavelet functions respectively, $j_0 \geq 0$, $\rho > 0$ plays a role similar to the inverse of bandwidth, and the summations involving k and ℓ are over $-\infty < k, \ell < \infty$.

As a prelude to constructing a wavelet estimator of γ, convert the dataset $\mathcal{Z}$ to a set of pairs $\widehat{W}_i = (\widehat{U}_i, \widehat{V}_i)$, for $1 \leq i \leq n$, where $\widehat{U}_i = \widehat{F}_X(X_i)$ and $\widehat{V}_i = \widehat{F}_Y(Y_i)$. In particular, $\widehat{U}_i$ is constructed by applying the transformation $\widehat{F}_X$, computed using the pooled marginal dataset $\mathcal{X}'$, to the ith value in the unpooled marginal dataset $\{X_1, \ldots, X_n\}$. Then, respective estimators of $a_{j_0 \ell k}$ and $b_{ij\ell k}$ are given by

$$\hat{a}_{j_0 \ell k} = \frac{1}{n} \sum_{r=1}^{n} \Phi_{j_0 \ell k}(\widehat{U}_r, \widehat{V}_r) \,, \quad \hat{b}_{ij\ell k} = \frac{1}{n} \sum_{r=1}^{n} \Psi_{ij\ell k}(\widehat{U}_r, \widehat{V}_r) \,,$$

and an estimator of γ is

$$\begin{aligned} \widehat{\gamma}(u, v) = & \sum_k \sum_\ell \hat{a}_{j_0 \ell k} \, \Phi_{j_0 \ell k}(u, v) \\ & + \sum_{i=1}^{3} \sum_{j=j_0}^{m} \sum_k \sum_\ell \hat{b}_{ij\ell k} \, I(|\hat{b}_{ij\ell k}| > 2\,\delta) \, \Psi_{ij\ell k}(u, v) \,, \end{aligned} \tag{2.3}$$

where

$$\delta = C_1 \, (n^{-1} \log n)^{1/2} \,, \tag{2.4}$$

denotes a threshold, m is a constant which should not exceed $C_2 \log n$ as n diverges, and $C_1, C_2 > 0$ are constants. This is a standard construction for a wavelet estimator; see, for example, Donoho et al. (1995). Practical implementation of $\widehat{\gamma}$ will be discussed in section 4.

If a method such as that above were employed to treat the quasi-parametric problem of Spiegelman and Park (2003), where models are available for the marginal distributions of X and Y but not for the joint distribution, then a technique such as maximum likelihood would be used to estimate the marginal distribution functions. The resulting estimators of F_X and F_Y would be substituted for the empirical distribution functions when computing $\widehat{U}_i = \widehat{F}_X(X_i)$ and $\widehat{V}_i = \widehat{F}_Y(Y_i)$.

3. THEORETICAL PROPERTIES

3.1. General issues. We begin by outlining some of the factors that influence performance. Since the additional data in $\mathcal{X}$ and $\mathcal{Y}$ relate only to marginal distributions, then, if the problem we are treating is intrinsically multivariate, first-order asymptotic properties, such as convergence rates and asymptotic variances, cannot be improved by incorporating the data in $\mathcal{X}$ and $\mathcal{Y}$, regardless of how large those samples might be. This result is readily proved using standard arguments.

On the other hand, if X and Y are independent then the problem of estimating f_{XY} is intrinsically univariate. In such cases the convergence rate of an estimator of f_{XY} can be improved from the two-dimensional rate to that for a single dimension. This can also occur if the copula density is so smooth that the problem of estimating it is of relatively low dimension.

Between the extremes of (a) approximate independence or substantial smoothness, where the copula-based method improves rates; and (b) non-independence or a non-smooth copula, where rate improvements are not obtained; there are many opportunities for enhancing performance in finite samples. The numerical work in section 4 will illustrate this point.

Arguably the most transparent way of capturing theoretically the potential for improved performance is to pass to the extreme case where both $\mathcal{X}$ and $\mathcal{Y}$ are infinite. There, the estimator at (2.1) has, in effect, the form

$$\hat{f}_{XY}(x,y) = f_X(x)\, f_Y(y)\, \widehat{\gamma}\{F_X(x), F_Y(y)\}\,. \tag{3.1}$$

We shall show in section 3.2 that in this setting the copula method leads to convergence-rate improvements in two classes of problems, where either (i) the copula density, γ, is close to its counterpart in the case of independence, i.e. to $\gamma \equiv 1$; or (ii) γ is particularly smooth.

We shall consider two "models" for γ, representing (i) and (ii) respectively. Let $0 \le \epsilon < \frac{1}{2}$ and define $\mathcal{S}_\epsilon$ to be the square $[\epsilon, 1-\epsilon]^2$. In model (i), γ varies with n and converges to the uniform density, in the sense that for δ defined in (2.4),

> $\gamma = 1 + g_n$, where the absolute values of the function g_n and its first derivatives are bounded in $\mathcal{S}_0$ by a constant multiple of δ, if $\epsilon = 0$, or in $\mathcal{S}_{\epsilon'}$ for some $\epsilon' \in (0, \epsilon)$, if $\epsilon > 0$. (3.2)

Of course, (3.2) implies that the marginals of the joint distribution of (X, Y) are asymptotically independent. In model (ii), γ is fixed and we ask that, for an integer $\nu \ge 1$,

> the function γ has ν bounded derivatives in $\mathcal{S}_0$, if $\epsilon = 0$, or in $\mathcal{S}_{\epsilon'}$ for some $\epsilon' \in (0, \epsilon)$, if $\epsilon > 0$. (3.3)

3.2. Theory for wavelet estimators of γ. We outline properties of $\widehat{\gamma}$ in the case where $\mathcal{X}$ and $\mathcal{Y}$ are of infinite size. Specifically, for $\nu \geq 1$ we assume that:

$$\begin{array}{l}\text{either (3.2) or (3.3) holds; } p = q = \infty\text{; } j_0 \text{ is fixed and } \rho = 1\text{; } \phi \text{ and}\\ \psi \text{ satisfy the usual standardisation conditions for wavelets of ``order''}\\ \nu\text{; the constant } C_1\text{, in the definition of the threshold at (2.4), satisfies}\\ C_1 > (2 \sup_{\mathcal{S}_\epsilon} \gamma)^{1/2}\text{; and } m = m(n)\text{, in (2.3), satisfies } 2^m \delta \to 0 \text{ and}\\ 2^{(2\nu+1)m} \delta \to \infty.\end{array} \tag{3.4}$$

The choice $j_0 = 0$, $\rho = 1$ is common in practice; it avoids the need to select ρ empirically, and results in only a logarithmic convergence-rate penalty.

When (3.2) holds, $\sup_{\mathcal{S}_\epsilon} \gamma$ in (3.4) may be interpreted as $\lim_{n\to\infty} \sup_{\mathcal{S}_\epsilon} \gamma$. In principle, $\sup_{\mathcal{S}_\epsilon} \gamma$ may be estimated from data, but simpler procedures are generally adequate; see section 4. The "usual standardisation conditions" referred to in (3.4) include, in particular, the assumption that $\int u^j\, \psi(u)\, du = 0$ for $0 \leq j \leq \nu - 1$, analogous to the condition defining the order of a conventional kernel function to be ν. They also include the condition that ϕ and ψ are bounded and compactly supported, and the wavelet expansion is orthonormal.

Assumption (3.4) implies the following performance bounds for $\widehat{\gamma}$. A proof of the theorem is similar to that of Proposition 2.1 of Hall and Patil (1995); an outline is given in the Appendix.

Theorem. *If* (3.4) *holds then*

$$\int_{\mathcal{S}_\epsilon} E(\widehat{\gamma} - \gamma)^2 = \begin{cases} O(\delta^2) & \text{under (3.2)} \\ O\big(\delta^{2\nu/(\nu+1)}\big) & \text{under (3.3).} \end{cases} \tag{3.5}$$

It follows that mean-square convergence rates arbitrarily close to n^{-1}, in a polynomial sense, can be achieved. To appreciate why, note that (3.5) implies that for any given $\xi > 0$, if g_n converges to zero sufficiently fast, assuming (3.2); or if ν is sufficiently large, assuming (3.3); then $\widehat{\gamma}$ converges to γ at mean-square rate $n^{\xi-1}$.

If ϵ were equal to 0, these results could not have been achieved using conventional kernel or orthogonal-series methods, since the performance of those techniques is hindered by discontinuities along the boundary, $\partial\mathcal{S}_0$, of $\mathcal{S}_0$. We would take $\epsilon > 0$ in (3.4) and (3.5) only when γ was unbounded at points of $\partial\mathcal{S}_0$; then, the left-hand side of (3.5) would not necessarily be finite unless $\epsilon > 0$.

By using wavelet functions ϕ and ψ that have sufficiently many derivatives, a Taylor-expansion argument may be used to extend the theorem to the general case where p and q increase at polynomial rates. In particular, it is not necessary to assume that either p or q is infinite. However, the numerical work reported in section 4 suggests that the assumption of sufficiently smooth ϕ and ψ is not necessary.

4. NUMERICAL PROPERTIES

4.1. Simulation study. In this section we describe a simulation study comparing the copula-based estimator $\hat{f}_{XY}$ with a standard bivariate kernel density estimator $\tilde{f}_{XY}$ which ignores the extra data.

All simulations were done with R (R Development Core Team, 2005). For implementation of the copula-based estimator we used the least asymmetric compactly supported wavelet function, or symmlet, s8, such that the mother wavelet ψ has four vanishing moments (see Daubechies, 1992). Then, ϕ has support $[0, S] = [0, 7]$ and the support of ψ is $[A, B] = [-3, 4]$. For computational efficiency of the wavelet coefficient estimators $\hat{a}_{j,\ell,k}$ it is important to utilise the property that for each j and each data point $\hat{U}_r$ there are at most S consecutive values of ℓ such that $2^j\,\hat{U}_r - \ell$ falls into the support of ϕ (see Herrick, Nason and Silverman, 2001). Analogous considerations are valid for $\hat{b}_{i,j,\ell,k}$. Using this fact, the estimated wavelet coefficients can be calculated directly from their definition. Alternatively, the pyramid algorithm can be implemented (see Vidakovic, 1999, p. 157).

In the simulations we used the empirical distribution functions $\widehat{F}_X$ and $\widehat{F}_Y$. Hence, the wavelet functions ϕ and ψ have only to be evaluated at points $2^j \frac{i}{n+p} - \ell$, where $i = 0, \ldots, n+p$, $j = j_0, \ldots, m$, $\ell = \lfloor 2^j \frac{i}{n+p} - S \rfloor, \ldots, \lceil 2^j \frac{i}{n+p} \rceil$ or $\ell = \lfloor 2^j \frac{i}{n+p} - B \rfloor, \ldots, \lceil 2^j \frac{i}{n+p} - A \rceil$, respectively, and at the corresponding points $2^j \frac{i}{n+q} - \ell$. Here, $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the largest integer $\leq x$, and the smallest integer $\geq x$, respectively.

Calculations were carried out using the Daubechies-Lagarias (1991, 1992) algorithm, via the implementation by Vidakovic. For real-data analysis it is recommended that continuous versions of $\widehat{F}_X$ and $\widehat{F}_Y$ be used, for example linear interpolations or the primitives of marginal density estimators, to obtain smoother distribution estimators. However, in simulations with many iterations this approach would be much more computationally demanding.

In each of 500 iterations the mean integrated squared error, or MISE, was approximated using 10,000 grid points on the square $[-4, 4]^2$. We set $\rho = 1$, $j_0 = 0$

and $m = \lfloor \frac{1}{2} \log n \rfloor$ in the definition of the copula density estimator $\hat{\gamma}$. No threshold was used when $j = 0$; we took $\delta = 2\,(n^{-1} \log n)^{1/2}$ for $j \geq 1$. Throughout we used $C_1 = 2$.

In Figure 1 results are displayed for different marginal distributions combined through a Gaussian copula with unit variances and correlation $\varrho \in \{0, 0.2, 0.4, 0.6\}$. The sample size for the paired observations is $n = 20$; we have no additional observations in the X sample, i.e. $p = 0$; and the number of extra Y observations varies, with $q \in \{20, 50, 100\}$. For Student's t marginals with 3 degrees of freedom, or Cauchy marginals, the copula-based estimator $\hat{f}_{XY}$ clearly outperforms the kernel estimator $\tilde{f}_{XY}$, even for $q = 20$ and even for high correlations. Indeed, this feature extends to $\varrho = 0.8$, although that value is not shown in the figure. When X is t_3 and Y standard normally distributed, the new estimator yields better results as long as the correlation is not too large, specifically $\varrho \in \{0, 0.2, 0.4\}$; for $\varrho \in \{0.6, 0.8\}$ the kernel estimator gives better results.

For the bivariate normal distribution, and when $\varrho = 0$, the new estimator outperforms the kernel density estimator even when $q = 20$. In the respective cases $\varrho = 0.2$ and $\varrho = 0.4$, $q = 50$ and $q = 100$ extra observations are needed to obtain better results for the new estimator. On the other hand, when $\varrho = 0.6$ the kernel estimator gives better results, even for very large q. Note, however, that we use the Gaussian reference bandwidth for the kernel estimator, so that for the Gaussian distribution, near-optimal results for the kernel estimator are to be expected. Nevertheless, when additional observations are also available in the X sample, the new copula-based estimator yields smaller MISE values than the kernel estimator; see the left-hand panel in Figure 3.

Figure 2 displays results for the same marginal distributions as Figure 1, but for the bounded Farlie-Gumbel-Morgenstern copula with density $\gamma(u, v) = 1 + a\,(2u - 1)\,(2v - 1)$. See Devroye (1986, p. 580) for generation of data from this copula. We took $a \in \{0, 0.2, 0.4, 0.6, 0.8\}$, where $a = 0$ corresponds to the independent case, and used $n = 20$ and $p = 0$. Except in the case of standard normal marginals and $q = 20$ extra observations, the new estimator clearly outperforms its kernel competitor. The right-hand panel of Figure 3 addresses the case of additional observations in the X sample.

In Figure 4 we consider two examples for sample size $n = 50$ and $p = q \in \{20, 50, 100, 150\}$ extra observations in both samples. In both panels the marginal distributions are t_3 and the copula is Gaussian in the left panel and Farlie-Gumbel-Morgenstern in the right.

To investigate how sensitive the method is with respect to choice of C_1, we considered $C_1 \in \{1.6, 2, 3, 4\}$ for the following cases: (i) Farlie-Gumbel-Morgenstern copula with parameter $a = 0.2$, standard normal marginals, $n = p = q = 20$; (ii) independent standard normal marginals, $n = 20$, $p = q = 50$; (iii) bivariate normal distribution with variances 1 and correlation $\varrho = 0.4$, $n = 20$, $p = q = 50$; and (iv) normal copula with correlation $\varrho = 0.4$ and t_3–distributed marginals, $n = p = q = 50$. In case (i) the smallest MISE, i.e. 2.327×10^{-4}, is obtained for $C_1 \in \{3, 4\}$; the values for $C_1 = 1.6$ and 2 are 2.36×10^{-4} and 2.33×10^{-4}, respectively. For cases (ii)–(iv) the MISE values hardly change for $C_1 \in \{2, 3, 4\}$, whereas for $C_1 = 1.6$, slightly larger values are observed; they are 2.12×10^{-4} compared to 2.04×10^{-4} in case (ii), 2.36×10^{-4} compared to 2.24×10^{-4} in case (iii), and 9.08×10^{-5} compared to 8.86×10^{-5} in case (iv). It can be concluded that the simulation results are not sensitive to the choice of C_1, provided it is not chosen too small.

4.2. Real-data example. Depending on seasonal demand, United Airlines operates two non-stop Los Angeles–Sydney flights, UA827 and UA839, both scheduled to arrive in the morning. Flight 827 operates only on Mondays, Wednesdays and Saturdays. On the other hand, 839 operates daily; it is scheduled to arrive at the same time each day, and about two hours after 827 when the latter operates. Occasionally, however, 827 is so late that it arrives after 839.

Below are two-vectors indicating the numbers of minutes late for flights 827 and 839 respectively, recorded on a sequence of Mondays, Wednesdays and Saturdays. A negative value indicates that that flight, on that day, arrived early, and a zero value indicates that the flight was right on time, to the nearest minute:

$(30, 4)$, $(865, 116)$, $(-1, 0)$, $(-5, 7)$, $(12, 13)$, $(10, 0)$, $(-5, 20)$, $(0, 15)$, $(32, 58)$, $(15, 85)$, $(30, 45)$, $(26, 30)$, $(6, 23)$, $(40, 55)$, $(3, 40)$, $(0, -8)$, $(11, 12)$, $(7, 13)$, $(-5, 9)$, $(-11, 6)$, $(-10, -20)$.

The numbers of minutes by which flight 839 was late, on Sundays, Tuesdays, Thursdays and Fridays during the same period, were:

20, 4, 5, 48, −30, −10, −22, −3, 80, −23, 0, 26, 10, 90, 90, 24, 30, 45, 17, 35, −10,−1, 30, 5, 18, 0, 40, 16, 6.

The data were recorded during part of the first quarter of 2005, and are in chronological order. However, owing to missing values, they are not always consecutive.

The upper panel of Figure 5 displays the copula-based estimator based on $n = 20$ paired observations (X, Y) and $q = 29$ extra Y observations, after deleting the outlier $(865, 116)$ and dividing the values by 60 to display fractions of hours. Here, we use linear interpolation to obtain continuous versions of the empirical distribution functions $\widehat{F}_X$ and $\widehat{F}_Y$. The lower panel of Figure 5 shows the kernel estimator using a bandwidth of 0.5 in each component.

The latter bandwidth is slightly smaller than either of the bandwidths recommended by the Gaussian reference method, but it is clear that the kernel estimator still produces a density which is too broadly supported, relative to the data. On the other hand, using a bandwidth for the kernel method which adequately reflects the spread of the data produces an estimator which is far too irregular. Resampling experiments show that, with high probability, the copula-based estimator based on the real dataset has smaller MISE than its kernel competitor. For these reasons, the copula-based method gives a more satisfactory estimator of the joint density of arrival times than does the standard kernel approach.

APPENDIX: Outline proof of theorem

To simplify notation, assume the supports of ϕ and ψ are both contained within the interval $[A, B]$, and define $I_j = \{2^j \epsilon - B, \ldots, 2^j(1-\epsilon) - A\}$. Let $\sum'$ denote summation over integers $i \in [1, 3]$, $j \in [j_0, m]$ and $k, \ell \in I_j$, and let $\sum''$ indicate summation over j, k, ℓ in the same ranges, with i held fixed at 1. Recall that $p = q = \infty$, and so $\widehat{F}_X = F_X$ and $\widehat{F}_Y = F_Y$. Then it may be proved that the left-hand side of (3.5) is dominated by $s_1 + s_2 + 2\,(s_3 + \ldots + s_6)$, where

$$
\begin{aligned}
s_1 &= \sum_{k \in I_{j_0}} \sum_{\ell \in I_{j_0}} E(\hat{a}_{j_0 \ell k} - a_{j_0 \ell k})^2\,, \quad s_2 = \sum_{i=1}^{3} \sum_{j>m} \sum_{k \in I_j} \sum_{\ell \in I_j} b_{ijk\ell}^2\,, \\
s_3 &= {\sum}' E\big(\hat{b}_{ijk\ell} - b_{ijk\ell}\big)^2\, I(|b_{ijk\ell}| > \delta)\,, \\
s_4 &= {\sum}' E\big\{\big(\hat{b}_{ijk\ell} - b_{ijk\ell}\big)^2\, I\big(\big|\hat{b}_{ijk\ell} - b_{ijk\ell}\big| > \delta\big)\big\}\,, \\
s_5 &= {\sum}' b_{ijk\ell}^2\, I(|b_{ijk\ell}| \le 4\delta)\,, \quad s_6 = {\sum}' b_{ijk\ell}^2\, P\big(\big|\hat{b}_{ijk\ell} - b_{ijk\ell}\big| > 2\delta\big)\,.
\end{aligned}
$$

To derive the theorem it suffices to show that each of $s_1, \ldots, s_6$ admits the bound on the right-hand side of (3.5).

It is relatively straightforward to prove that $s_1 = O(n^{-1})$. Employing a Taylor expansion of γ up to terms of νth order, and utilising the fact that the first $\nu - 1$ moments of the mother wavelet vanish, it can be proved that $|b_{ijk\ell}|$ is bounded by a constant multiple of $2^{-(1+\nu)j}$ if (3.3) holds, and of $\delta\, 2^{-2j}$ under (3.2), uniformly in $(k,\ell) \in I_j \times I_j$; call this property (P). Now, (P) leads to the bounds, $s_2 = O(2^{-2\nu m}) = O(\delta^{2\nu/(\nu+1)})$ under (3.3), and $s_2 = O(\delta^2)$ if (3.2) holds. Also, (P) and the fact that $E(\hat{b}_{ijk\ell} - b_{ijk\ell})^2 \le n^{-1} \sup_{\mathcal{S}_\epsilon} \gamma$ imply that $s_3 = O(m/n) = O(\delta^2)$ under (3.2), and $s_3 = O(mn^{-1}\,\delta^{-2/(1+\nu)}) = O(\delta^{2\nu/(\nu+1)})$ if (3.3) is valid.

For $r = 4,5,6$, let s_{ri} denote the version of s_r obtained if we fix i, and in particular do not sum over i in the definition of s_{ri}. We shall derive bounds for s_r in the case of s_{r1}. Other components s_{ri} can be handled similarly.

Let $u \ge 2$ be an integer, and put $v^{-1} = 1 - u^{-1}$. Then, $E\{(\hat{b}_{ijk\ell} - b_{ijk\ell})^{2u}\} = O(n^{-u})$, uniformly in j,k,ℓ. Therefore, by Hölder's and Bernstein's inequalities,

$$\begin{aligned} s_{41} &= O\bigg[\frac{1}{n}\sum{}'' \left\{P\big(|\hat{b}_{1jk\ell} - b_{1jk\ell}| > \delta\big)\right\}^{1/v}\bigg] \\ &= O\bigg(\frac{1}{n}\sum{}'' \exp\bigg[-\frac{(n\delta)^2}{2v}\Big\{n \sup_{\mathcal{S}_\epsilon} \gamma + \big(2^j/3\big)\, n\delta \sup|\phi| \sup|\psi|\Big\}^{-1}\bigg]\bigg) \\ &= O\bigg(\frac{1}{n}\sum{}'' \exp\bigg[-\frac{\log n}{2v}\Big\{C_1^{-2} \sup_{\mathcal{S}_\epsilon} \gamma + O\big(2^m\delta\big)\Big\}^{-1}\bigg]\bigg) = O\big(4^m\, n^{-1-(\kappa/v)}\big)\,, \end{aligned} \tag{A.1}$$

where C_1 denotes the constant in the threshold at (2.4), and $\kappa > 1$ does not depend on v. If $v > 1$ is sufficiently close to 1, or equivalently, if $u \ge 2$ is sufficiently large, then $4^m\, n^{-1-(\kappa/v)} = O(\delta^2)$ under (3.2), and equals $O(\delta^{2\nu/(\nu+1)})$ if (3.3) holds. Result (A.1) implies that the same bounds apply to s_{41}. A similar argument, although starting from the bound $s_{61} = O\{\sum'' w_j^2\, P(|\hat{b}_{1jk\ell} - b_{1jk\ell}| > 2\delta)\}$ where, using property (P), $w_j = 2^{-(1+\nu)j}$ if (3.2) holds and equals $\delta\, 2^{-2j}$ under (3.3), shows that s_{61} also enjoys the bound on the right-hand side of (3.5).

Using (P) again we deduce that if (3.3) holds, $b_{1jk\ell}^2\, I(|b_{1jk\ell}| \le 4\delta)$ is bounded above by a constant multiple of $\min(\delta, 2^{-(1+\nu)j})^2$. Adding this quantity over the indices j,k,ℓ, in the respective ranges indicated by $\sum''$, we deduce that s_{51} is bounded above by a constant multiple of $\delta^{2\nu/(\nu+1)}$. A similar argument in the case where (3.2) holds produces the bound $s_{51} \le \text{const.}\, \delta^2$. This completes the proof.

REFERENCES

DAUBECHIES, I. (1992). *Ten lectures on wavelets.* CBMS-NSF Regional Conference Series in Applied Mathematics, 61., Society for Industrial and Applied

Mathematics (SIAM), Philadelphia.

DAUBECHIES, I. AND LAGARIAS, J. (1991). Two-scale difference equations I. Existence and global regularity of solutions, *SIAM J. Math. Anal.* **22**, 1388–1410.

DAUBECHIES, I. AND LAGARIAS, J. (1992). Two-scale difference equations II. Local regularity, infinite products of matrices and fractals, *SIAM J. Math. Anal.* **23**, 1031–1079.

DEVROYE, L. (1986). *Non-Uniform Random Variate Generation.* Springer-Verlag, New York.

DONOHO, D., JOHNSTONE, I.M., KERKYACHARIAN, G. AND PICARD, D. (1995). Wavelet shrinkage: asymptopia? (With discussion.) *J. Roy. Statist. Soc.* Ser. B **57**, 301–369.

FERMANIAN, J.-D. AND SCAILLET, O. (2002). Nonparametric estimation of copulas for time series. FAME Research Paper No. 57. `http://ssrn.com/abstract=372142`

GIJBELS, I. AND MIELNICZUK, J. (1990). Estimating the density of a copula function. *Comm. Statist. Theory Methods* **19**, 445–464.

HALL, P. AND PATIL, P. (1995). Formulae for mean integrated squared error of wavelet-based density estimators. *Ann. Statist.* **23**, 905–928.

HERRICK, D.R.M., NASON, G.P. AND SILVERMAN, B.W. (2001). Some new methods for wavelet density estimation. *Sankhyā Ser. A* **63**, 394–411.

JONES, M.C. (1993). Kernel density estimation when the bandwidth is large. *Austral. J. Statist.* **35**, 319–326.

LIEBSCHER, E. (2005). Semiparametric density estimators using copulas. *Comm. Statist. Theory Methods* **34**, 59–71.

OLKIN, I. AND SPIEGELMAN, C.H. (1987). A semiparametric approach to density estimation. *J. Amer. Statist. Assoc.* **82**, 858–865.

R DEVELOPMENT CORE TEAM (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`

SANCETTA, A. (2003). Nonparametric estimation of multivariate distributions with given marginals.: L_2 theory. Cambridge Working Papers in Economics No. 0320.

SANCETTA, A. AND SATCHELL, S. (2004). The Bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric Theory* **20**, 535–562.

SCHUSTER, E. AND YAKOWITZ, S. (1979). Contributions to the theory of nonparametric regression, with application to system identification. *Ann. Statist.* **7**, 139–149.

SPIEGELMAN, C.H. AND PARK, E.S. (2003). Nearly nonparametric multivariate density estimates that incorporate marginal parametric density information. *Amer. Statist.* **57**, 183–188.

VIDAKOVIC, B. (1999). *Statistical Modeling by Wavelets.* Wiley, New York.

VIDAKOVIC, B. Lagarias-Daubechies' algorithm in action – Splus implementation. `www.isye.gatech.edu/~brani/.publichtml/Wiley/soft/LD.pdf`