Journal of Nonparametric Statistics
Vol. 00, No. 00, Month 2011, 1–23

RESEARCH ARTICLE

Optimal plug-in estimators for multivariate distributions with conditionally independent components

Ursula U. Müller[a*] , Anton Schick[b] and Wolfgang Wefelmeyer[c]

[a]*Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA*; [b]*Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902-6000, USA*; [c]*Mathematical Institute, University of Cologne, 50931 Cologne, Germany*

(*Received ; final version received*)

The usual estimator for the expectation of a function of a random vector is the empirical estimator. Assume that some of the components of the random vector are conditionally independent given the other components. We construct a plug-in estimator for the expectation that uses this information, prove a central limit theorem for the estimator, and show that the estimator is asymptotically efficient in the sense of a nonparametric version of the convolution theorem of Hájek and Le Cam.

Keywords: smoothed empirical estimator, multivariate density estimator, local asymptotic normality

2010 AMS Subject Classifications: Primary: 62G05; Secondary: 62G20

1. Introduction

We want to estimate an expectation $E[f(X)]$ from independent copies $X_1, \dots, X_n$ of X. The usual estimator is the empirical estimator $\mathbb{E}f = (1/n)\sum_{j=1}^n f(X_j)$. If we can decompose X_j into d independent components $X_j = (X_{1j}, \dots, X_{dj})$, then an improved estimator of $E[f(X)]$ is the von Mises statistic

$$\mathbb{M} = \frac{1}{n^d} \sum_{j_1=1}^n \cdots \sum_{j_d=1}^n f(X_{1j_1}, \cdots, X_{dj_d}).$$

*Corresponding author. Email: uschi@stat.tamu.edu

ISSN: 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/1048525YYxxxxxxxx
http://www.informaworld.com

This estimator is the empirical estimator

$$\int \dots \int f(x_1, \dots, x_d) \mathbb{P}_1(dx_1) \cdots \mathbb{P}_d(dx_d)$$

based on the marginal empirical distributions $\mathbb{P}_i(A) = (1/n) \sum_{j=1}^n \mathbf{1}(X_{ij} \in A)$ for $i = 1, \dots, d$. It makes use of the independence of the components of X_j and has a smaller variance than the empirical estimator $\mathbb{E}f$. The von Mises statistic (or V-statistic) can be viewed as a special case of a generalized U-statistic; see e.g. Serfling (1980, sec. 5.1.2 and sec. 5.1.3.).

In this paper we consider an intermediate estimation problem that has not been treated in the literature before: We do not have independent components, but some of the components are conditionally independent given the other components.

Two potential applications show why this problem is of interest. Firstly, we can have conditional independence in *regression models*. Suppose the outcome is Y, and the covariates are X and Z. Assume that the covariate X is *sufficient for predicting* Y in the sense that the conditional distribution of Y given (X, Z) does not depend on Z. See Causeur and Dhorne (2003) for maximum likelihood estimation in this model. The model is equivalent to the conditional independence of Y and Z given X.

Another application are *graphical models*. Then the joint density of a random vector $(X_1, \dots, X_d)$ is expressed as a product of the one-dimensional marginal density of the first component and $d-1$ one-dimensional conditional densities of the remaining components of the vector given the first component. We refer to Lauritzen (1996) or Pearl (2000) for graphical models and causal inference.

The problem of *testing* for conditional independence has been treated before. The techniques are different. We refer to unpublished papers by Linton and Gozalo (1999) and Fernandes and Flôres (1999), and to Delgado and González Manteiga (2001), Delgado, Domínguez and Lavergne (2006), Su and White (2007, 2008), and Huang (2010).

For notational simplicity we will restrict attention to the simplest situation, in which we assume that $W_1, \dots, W_n$ are independent copies of a three-dimensional random vector $W = (X, Y, Z)$, and Y and Z are conditionally independent given X. How can we use this information for estimating an expectation $E[f(W)]$? We assume that W has a density $p(x, y, z)$. Then X, (X, Y) and (X, Z) also have densities, say $m(x)$, $s(x, y)$ and $t(x, z)$, respectively, and the conditional densities of Y given X and of Z given X are

$$q(y|x) = \frac{s(x, y)}{m(x)}, \qquad r(z|x) = \frac{t(x, z)}{m(x)}, \tag{1.1}$$

respectively. The conditional independence of Y and Z given X means that the conditional density of (Y, Z) given X factors as $c(y, z|x) = q(y|x)r(z|x)$. We can therefore express the expectation $E[f(W)]$ as

$$\begin{aligned} E[f(W)] &= \iiint f(x, y, z) q(y|x) r(z|x) m(x) \, dy \, dz \, dx \\ &= \iiint f(x, y, z) \frac{s(x, y) t(x, z)}{m(x)} \, dx \, dy \, dz. \end{aligned}$$

We estimate $E[f(W)]$ by plugging kernel estimators $\hat{m}$, $\hat{s}$, $\hat{t}$ for m, s, t into this expression.

Our main result is Theorem 2.1. It shows that the plug-in estimator

$$\mathbb{T} = \iiint f(x,y,z)\frac{\hat{s}(x,y)\hat{t}(x,z)}{\hat{m}(x)}\,dx\,dy\,dz \tag{1.2}$$

converges at the rate $n^{-1/2}$ and is asymptotically normal. In Section 3 we prove that our estimator is also asymptotically efficient. The asymptotic distribution of our estimator will be unchanged if we use spline estimators or series estimators in place of the kernel estimators; see Masri and Redner (2005) and Efromovich (1999) for convergence results on such estimators. One could also use two Nadaraya–Watson estimators for conditional expectations, or estimate the conditional densities $q(y|x)$ and $r(z|x)$ directly; see Efromovich (2005, 2010) for such estimators.

If Y and Z are known to be independent, not just conditionally independent given X, and $f(x,y,z)$ does not depend on x, i.e., $f(x,y,z) = g(y,z)$, then an efficient estimator of $E[g(Y,Z)]$ is the von Mises statistic $(1/n^2)\sum_{i=1}^n\sum_{j=1}^n g(Y_i,Z_j)$. An example is $P(Y < Z)$. Suppose now that $f(x,y,z)$ also depends on x. If Y and Z are again independent, with densities φ and ϕ, and $\rho(\cdot|y,z)$ denotes the conditional density of X given $(Y,Z) = (y,z)$, then we can write $E[f(X,Y,Z)] = \iiint f(x,y,z)\varphi(y)\phi(z)\rho(x|y,z)\,dx\,dy\,dz$ and obtain an efficient estimator of $E[f(X,Y,Z)]$ by plugging in estimators for φ, ϕ and ρ. The proof would be similar to that for our estimator $\mathbb{T}$ given above.

In order to describe the asymptotic variance of our estimator, we write conditional expectations as

$$\begin{aligned}
Qf(X,Z) &= E(f(X,Y,Z)|X,Z) = \int f(X,y,Z)q(y|X)\,dy,\\
Rf(X,Y) &= E(f(X,Y,Z)|X,Y) = \int f(X,Y,z)r(z|X)\,dz,\\
RQf(X) &= E(f(X,Y,Z)|X) = \iint f(X,y,z)q(y|X)r(z|X)\,dy\,dz.
\end{aligned}$$

Our Theorem 2.1 implies that the estimator $\mathbb{T}$ obeys the expansion

$$\mathbb{T} = \frac{1}{n}\sum_{j=1}^n \big(Rf(X_j,Y_j) + Qf(X_j,Z_j) - RQf(X_j)\big) + o_p(n^{-1/2}).$$

By the central limit theorem, the standardized estimator $n^{1/2}(\mathbb{T} - E[f(W)])$ converges to a centered normal distribution. The variance of this distribution is called the *asymptotic variance* of $\mathbb{T}$; it is equal to

$$\begin{aligned}
&E\big[\big(Rf(X,Y) + Qf(X,Z) - RQf(X) - E[f(W)]\big)^2\big]\\
&\quad = E\big[\big(Rf(X,Y) - RQf(X)\big)^2\big] + E\big[\big(Qf(X,Z) - RQf(X)\big)^2\big] + \mathrm{Var}(RQf(X)).
\end{aligned}$$

It is smaller than the asymptotic variance $E[(f(W) - E[f(W)])^2]$ of the empirical estimator by the amount

$$E\big[\big(f(X,Y,Z) - Rf(X,Y) - Qf(X,Z) + RQf(X)\big)^2\big],$$

which is nonzero except when $f(x,y,z)$ does not depend on both y and z. We shall now show, by means of two examples, that this amount can be a substantial fraction of asymptotic variance of the empirical estimator.

Let us first consider estimation of $P(Y<Z)$. Our estimator is

$$\mathbb{T}=\int_\alpha^\beta \iint\limits_{y<z} \frac{\hat{s}(x,y)\hat{t}(x,z)}{\hat{m}(x)}\,dy\,dz\,dx,$$

where $[\alpha,\beta]$ is the support of X. Here we have $Qf(X,Z)=F(Z|X)$ and $Rf(X,Y)=1-G(Y|X)$ with $G(y|x)=P(Z\le y|X=x)$ and $F(z|x)=P(Y\le z|X=x)$, and $RQf(X)=P(Y<Z|X)$. In particular, if the conditional distribution functions F and G are identical, then conditionally given X, both $1-G(Y|X)$ and $F(Z|X)$ have a uniform distribution and are independent. Moreover, we have the identity $P(Y<Z)=P(Y<Z|X)=1/2$, and we find that our estimator has asymptotic variance $1/6$. In contrast, the asymptotic variance of the empirical estimator is $1/4$. The reduction amount is $1/12$ and equals $1/3$ of the asymptotic variance of the empirical estimator.

Next we look at estimating $H(y,z)=P(Y\le y,Z\le z)$ for fixed reals y and z. The asymptotic variance of the empirical estimator is $H(y,z)(1-H(y,z))$. We have $Rf(X,Y)=\mathbf{1}[Y\le y]G(z|X)$, $Qf(X,Z)=F(y|X)\mathbf{1}[Z\le z]$ and $RQf(X)=F(y|X)G(z|X)$. The reduction amount simplifies to $E[(F(y|X)-F^2(y|X))(G(z|X)-G^2(z|X))]$. If $F(y|X)=D(y/u(X))$ and $G(z|X)=D(z/v(X))$ for a distribution function D and positive functions u and v, then the reduction amount for $y=z=0$ becomes $(D(0)-D^2(0))^2$. For $D(0)=1/2$ the reduction amount is $1/16$. This is $1/3$ of the asymptotic variance of the empirical estimator. The latter is $3/16$ as $H(0,0)$ equals $1/4$.

In the following table we list values of the relative reduction amount

$$\varrho(y,z)=\frac{E[(F(y|X)-F^2(y|X))(G(z|X)-G^2(z|X))]}{H(y,z)-H^2(y,z)}$$

for selected values of y and z with $F(y|X)=\Phi(y-X)$ and $G(z|X)=\Phi(z-X^2)$, and with X having a uniform distribution on $[-1,1]$. Here Φ denotes the standard normal distribution function. These values were obtained by numerical integration using the function `integrate` in `R`. As can be seen, the improvements can be astonishingly large.

Table 1. Relative reduction amounts $\varrho(x,y)$

$y\backslash z$	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5
-1.5	0.83	0.77	0.67	0.54	0.38	0.24	0.13
-1.0	0.73	0.67	0.59	0.47	0.33	0.21	0.11
-0.5	0.59	0.55	0.48	0.39	0.28	0.18	0.09
0.0	0.43	0.41	0.37	0.31	0.24	0.16	0.09
0.5	0.29	0.27	0.25	0.22	0.18	0.13	0.08
1.0	0.16	0.16	0.15	0.14	0.13	0.10	0.07
1.5	0.08	0.08	0.08	0.08	0.08	0.07	0.06

Simulations in Section 4 for these two examples show that even for a rather small sample size such as $n=30$, our estimator substantially improves on the empirical estimator. Furthermore, our estimators are fairly insensitive to the choice of bandwidth. The

simulated relative reduction amounts of the mean squared error of our estimator in the second example, over that of the empirical estimator, can even be considerably better than those suggested by the theoretical values in Table 1.

The proofs of our results are carried out under the assumption that the density m is quasi-uniform on some compact interval $[\alpha, \beta]$, which means that m is bounded and bounded away from zero on that interval and vanishes outside the interval. This requires the use of boundary-adjusted kernel estimators. We assume that s and t have continuous second derivatives with certain moment conditions.

Often we do not know the boundary points α and β. They can however be estimated at a good rate. We therefore expect the correspondingly adjusted estimator for $E[f(W)]$ to behave like $\mathbb{T}$.

There is a large literature on "plug-in estimators", i.e., estimators that involve integrating a density estimator. Our estimator $\mathbb{T}$ integrates over more than one density estimator, and one of them is in the denominator. This makes the proof more involved. The proof also requires convergence of the density estimators in *weighted* L_1-norms. Such results have been obtained before, but not for boundary-adjusted density estimators. We refer to Schick and Wefelmeyer (2004, 2007) and to Müller, Schick and Wefelmeyer (2005).

Our method also shows that under the assumption of conditional independence of Y and Z given X, an expectation $E[f(X, Y, Z)]$ can be estimated even if only pairs (X, Y) and (X, Z) are observed, for example

$$(X_1, Y_1), \ldots, (X_{n_1}, Y_{n_1}) \quad \text{and} \quad (X_{n_1+1}, Z_{n_1+1}), \ldots, (X_{n_1+n_2}, Z_{n_1+n_2}).$$

Since in this case no observations on the triple (X, Y, Z) are available, the usual empirical estimator cannot be computed for functions $f(X, Y, Z)$ depending on both Y and Z.

Our paper is organized as follows. In Section 2 we introduce the boundary-adjusted kernel density estimators for m, s and t used in the plug-in estimator. Their asymptotic behavior in suitable norms is described in two lemmas in Section 5. The asymptotic distribution of our plug-in estimator is given in Theorem 2.1. In Section 3 we prove that the plug-in estimator is asymptotically efficient. The results of a simulation study are discussed in Section 4. Section 5 contains the aforementioned lemmas and the proof of Theorem 2.1.

2. Stochastic expansion of the plug-in estimator

Let $W = (X, Y, Z)$ have distribution P. Suppose that Y and Z are conditionally independent given X. Then the conditional distribution of (Y, Z) given X factors into the conditional distributions of Y given X, and Z given X,

$$C(dy, dz|X) = Q(dy|X)R(dz|X).$$

With M denoting the distribution of X, we have

$$P(dx, dy, dz) = M(dx)Q(dy|x)R(dz|x).$$

We want to estimate an expectation $E[f(W)]$ for a bounded function f. Assume that X, Y, Z are real-valued, and that W has a density p. Then the distributions M, S, T of X, (X, Y), (X, Z) have densities m, s, t, respectively, and the conditional distributions

Q, R have densities $q(y|x)$ and $r(z|x)$ as in (1.1), and we can estimate $E[f(W)]$ by the plug-in estimator (1.2).

We assume for simplicity that M is supported on a finite interval $[\alpha, \beta]$ and has a density that is bounded away from zero on that interval. This will require modifications of the density estimators near the boundary. The usual estimator would be

$$\tilde{m}(x) = \frac{1}{n}\sum_{j=1}^{n} k_b(x - X_j)$$

with $k_b(x) = k(x/b)/b$ for a bounded symmetric kernel k with support $[-1, 1]$ and a bandwidth b. This estimator is not suitable near the endpoints α and β. The literature on boundary-adjusted density estimators is large. Here, however, we need convergence in a weighted L_1-norm, for which results are not available. For this reason we will state and prove the required convergence results in two lemmas in Section 5, see Lemmas 5.1 and 5.2. They are perhaps not of separate interest. So we will consider only the simplest choice. We use $\tilde{m}$ only on the interval $[\alpha + b, \beta - b]$, take kernel estimators with one-sided kernels at the endpoints α and β, and interpolate linearly on the two remaining intervals $(\alpha, \alpha + b)$ and $(\beta - b, \beta)$. The resulting estimator can be written

$$\hat{m}(x) = \begin{cases} \tilde{m}_\alpha + \frac{x-\alpha}{b}(\tilde{m}(\alpha + b) - \tilde{m}_\alpha), & \alpha \le x < \alpha + b, \\ \tilde{m}(x), & \alpha + b \le x \le \beta - b, \\ \tilde{m}(\beta - b) + \frac{x-\beta+b}{b}(\tilde{m}_\beta - \tilde{m}(\beta - b)), & \beta - b < x \le \beta, \end{cases}$$

where

$$\tilde{m}_\alpha = \frac{1}{n}\sum_{j=1}^{n} h_b(X_j - \alpha) \quad \text{and} \quad \tilde{m}_\beta = \frac{1}{n}\sum_{j=1}^{n} h_b(\beta - X_j),$$

with $h_b(x) = h(x/b)/b$ and h a bounded kernel with support $[0, 1]$ and $\int xh(x)\,dx = 0$.

We can write

$$\hat{m}(x) = \frac{1}{n}\sum_{j=1}^{n} K_b(x, X_j)$$

with

$$K_b(x, y) = \begin{cases} \big(1 - \frac{x-\alpha}{b}\big)h_b(y - \alpha) + \frac{x-\alpha}{b}k_b(\alpha + b - y), & \alpha \le x < \alpha + b, \\ k_b(x - y), & \alpha + b \le x \le \beta - b, \\ \frac{\beta-x}{b}k_b(\beta - b - y) + \big(1 - \frac{\beta-x}{b}\big)h_b(\beta - y), & \beta - b < x \le \beta. \end{cases}$$

Similarly, for a bandwidth c and $x \in [\alpha, \beta]$ we estimate $s(x, y)$ by

$$\hat{s}(x, y) = \frac{1}{n}\sum_{j=1}^{n} K_b(x, X_j)k_c(y - Y_j),$$

and $t(x,z)$ by

$$\hat{t}(x,z) = \frac{1}{n}\sum_{j=1}^{n} K_b(x,X_j)k_c(z-Z_j).$$

For other boundary adjustments of density estimators we refer to Müller (1993), Müller and Stadtmüller (1999), Masri and Redner (2005), Efromovich (2010), Marshall and Hazelton (2010), and Bouezmarni and Rombouts (2010).

For $\eta \geq 0$, we let H_η denote the set of measurable functions ψ on the strip $S = [\alpha,\beta]\times\mathbb{R}$ which satisfy

$$\|\psi\|_{H_\eta} = \sup_{\alpha\leq x\leq\beta}\int (1+|y|)^\eta |\psi(x,y)|\,dy < \infty.$$

We now state our main result, Theorem 2.1. The assumptions are explained subsequently in Remarks 1 and 2. It is straightforward to show that conditions (2.1)–(2.4) are satisfied for smooth f; they also cover discontinuous f.

Theorem 2.1 : *Assume that the density m vanishes outside a finite interval $[\alpha,\beta]$ and is positive on $[\alpha,\beta]$. Suppose the densities s and t are twice continuously differentiable on S with gradients $\dot{s}$ and $\dot{t}$ and Hessians $\ddot{s}$ and $\ddot{t}$, respectively, and that s, t, $\|\ddot{s}\|^2$ and $\|\ddot{t}\|^2$ belong to H_η for some $\eta > 1$. Let f be bounded by 1 and*

$$E[(f(X+u,Y+v,Z+w)-f(X,Y,Z))^2] \to 0 \quad \text{as } u,v,w \to 0 \tag{2.1}$$

and, for $I_b = \mathbf{1}[\alpha+2b \leq X \leq \beta-2b]$ and some $\xi \in (2,3]$,

$$E\Big[I_b \iint Rf(X+bu,Y+cv)k(u)k(v)\,du\,dv\Big] = E[I_bRf(X,Y)] + O(b^\xi+c^\xi), \tag{2.2}$$

$$E\Big[I_b \iint Qf(X+bu,Z+cv)k(u)k(v)\,du\,dv\Big] = E[I_bQf(X,Y)] + O(b^\xi+c^\xi), \tag{2.3}$$

$$E\Big[I_b \iint RQf(X+bu)k(u)\,du\Big] = E[I_bRQf(X)] + O(b^\xi). \tag{2.4}$$

Let the bandwidths b and c satisfy $nb^2c^2 \to \infty$, $nb^{2\xi} \to 0$, and $nc^{2\xi} \to 0$. Then

$$\mathbb{T} = \frac{1}{n}\sum_{j=1}^{n}\big(Rf(X_j,Y_j)+Qf(X_j,Z_j)-RQf(X_j)\big) + o_p(n^{-1/2}).$$

Remark 1: If the assumptions of Theorem 2.1 hold with $\xi = 3$, then the conclusion of the theorem holds with $c \sim b \sim n^{-1/5}$. For simplicity in notation we have used the same bandwidth c for estimating s and t. Our results continue to hold if we use different bandwidths, as long as these bandwidths are proportional to c with c as in the theorem. We should also point out that the assumptions on s and t imply that the moments $E[|Y|^\eta]$ and $E[|Z|^\eta]$ are finite for the η in the theorem.

Remark 2: The conditions (2.2)–(2.4) typically require the use of higher-order kernels. For example, consider (2.2). Assume that the functions

$$F_b(u,v) = E[I_b Rf(X+u, Y+v)] = E[I_b f(X+u, Y+v, Z)], \quad |u| \le b, |v| \le c,$$

are twice differentiable and their derivatives of order 2 are Lipschitz, uniformly in b and c for small positive b and c. Then the use of a kernel of order 3 yields (2.2) with $\xi = 3$. (By kernel of order 3 we mean a (signed) kernel whose first two moments vanish and whose third moment is non-zero.)

Example 2.2 Our plug-in estimator of $P(Y < Z)$ is

$$\mathbb{T} = \int_\alpha^\beta \iint\limits_{y<z} \frac{\hat{s}(x,y)\hat{t}(x,z)}{\hat{m}(x)}\, dy\, dz\, dx.$$

In this case we have $f(X,Y,Z) = \mathbf{1}[Y < Z]$, $Rf(y,z) = 1 - G(y|x) = P(Z > y|X = x)$, $Qf(x,z) = F(z|x) = P(Y < z|X = x)$ and $RQf(x) = P(Y < Z|X = x) = H(x)$. Let m, s and t satisfy the assumptions of Theorem 2.1 and assume furthermore that F and G have bounded continuous partial derivatives of order three and that H has a bounded third derivative. Then (2.2)–(2.4) hold with $\xi = 3$, provided k is of order 3, while (2.1) holds trivially. Thus Theorem 2.1 yields the stochastic expansion

$$\mathbb{T} = \frac{1}{n}\sum_{j=1}^n \big(P(Y_j < Z_j|X_j, Y_j) + P(Y_j < Z_j|X_j, Z_j) - P(Y_j < Z_j|X_j)\big) + o_p(n^{-1/2}).$$

Example 2.3 Our plug-in estimator of the joint distribution function

$$H(y_0, z_0) = P(Y \le y_0, Z \le z_0)$$

of Y and Z at a fixed point (y_0, z_0) is

$$\mathbb{T} = \int_\alpha^\beta \int_{-\infty}^{y_0} \int_{-\infty}^{z_0} \frac{\hat{s}(x,y)\hat{t}(x,z)}{\hat{m}(x)}\, dz\, dy\, dx.$$

Here we have $Rf(x,y) = \mathbf{1}[y \le y_0]G(z_0|x)$, $Qf(x,z) = F(y_0|x)\mathbf{1}[z \le z_0]$, and $RQf(x) = F(y_0|x)G(z_0|x)$. Let m, s and t satisfy the assumptions of Theorem 2.1 and assume furthermore that F and G have bounded continuous partial derivatives of order three. Then one derives conditions (2.2)–(2.4) with $\xi = 3$ and k a kernel of order 3, while (2.1) holds trivially. Thus Theorem 2.1 yields the stochastic expansion

$$\mathbb{T} = \frac{1}{n}\sum_{j=1}^n \mathbf{1}[Y_j \le y_0]G(z_0|X_j) + F(y_0|X_j)\mathbf{1}[Z_j \le z_0] - F(y_0|X_j)G(z_0|X_j) + o_p(n^{-1/2}).$$

3. Efficiency of the plug-in estimator

For parametric models, Hájek (1970) and Le Cam (1972) introduced an asymptotic efficiency concept for (regular) estimators of one-dimensional (differentiable) functionals on

locally asymptotically normal families. The concept generalizes to infinite-dimensional parameter spaces, using the observation of Stein (1956) that it suffices to look at the one-dimensional submodel that is *least favorable* for the functional, in the sense that the achievable lower variance bound is largest among all one-dimensional submodels.

In particular, Koshevnik and Levit (1976) show that the empirical estimator $\mathbb{E}f = (1/n)\sum f(W_i)$ is efficient for $E[f(W)]$ if the model is *nonparametric*, i.e., if no structural assumptions are made on the distribution P of W. We briefly recall their result. Fix P. The asymptotic efficiency concept depends only on first-order approximations, at P, of the model and of the functional $E[f(W)] = \int f(w)P(dw)$ that we want to estimate. The one-dimensional submodels may be described by their likelihood ratios with respect to the fixed P. For the nonparametric model, these are of the form $P_{tc}(dw) = P(dw)(1+tc(w))$ with t running through $\mathbb{R}$, where c is any function that fulfills $E[c(W)] = 0$. The latter condition is needed for P_{tc} to be a probability distribution. We may and will assume for convenience, and without loss of generality, that c is bounded. The functions c are called *local parameters* at P. They form the *tangent space* of the model at P. For the nonparametric model, the tangent space is dense in

$$L_{2,0}(P) = \{c \in L_2(P) : E[c(W)] = 0\}.$$

It suffices to consider $t = n^{-1/2}$. Then we write P_{nc} for $P_{n^{-1/2}c}$. For the joint law of n observations $(W_1, \ldots, W_n)$ under P and P_{nc} we write P^n and P^n_{nc}, respectively. A Taylor expansion shows that we have *local asymptotic normality* at P^n,

$$\log \frac{dP^n_{nc}}{dP^n} = n^{-1/2}\sum_{j=1}^{n} c(W_j) - \frac{1}{2}E[c^2(W)] + o_p(n^{-1/2}).$$

Proofs of local asymptotic normality under minimal assumptions are in Le Cam (1956, 1966, 1969); see also Bickel, Klaassen, Ritov and Wellner (1998, sec. 2.1, Proposition 2) for a version that is uniform in the parameter.

The squared norm $E[c^2(W)]$ induces an inner product $(c, c') = E[c(W)c'(W)]$ on the closure $L_{2,0}(P)$ of the tangent space. The least favorable one-dimensional submodel for $E[f(W)]$ is given by the *gradient* of $E[f(W)]$, i.e. the direction of steepest ascent, in terms of this inner product. This is the function $k \in L_{2,0}(P)$ such that

$$n^{1/2}\Big(\int f(w)P_{nc}(dw) - \int f(w)P(dw)\Big) \to (k, c) \quad \text{for all} \quad c \in L_{2,0}(P).$$

It is easy to see that $k = f - E[f(W)]$. We will now show that the variance of the gradient is a lower variance bound for "regular" estimators of $E[f(W)]$. An estimator $\hat{\kappa}$ of $E[f(W)]$ is called *regular* at P with *limit* L if L is a random variable such that

$$n^{1/2}(\hat{\kappa} - E[f(W)]) \Rightarrow L \quad \text{under } P_{nc} \quad \text{for all} \quad c \in L_{2,0}(P).$$

The convolution theorem of Hájek and Le Cam says that L is distributed as a convolution $M + \sigma N$, where M and N are independent and N is standard normal, and $\sigma^2 = E[k^2(W)] = \mathrm{Var} f(W)$. The random variable σN is more concentrated in symmetric intervals than $M + \sigma N$. This justifies calling $\hat{\kappa}$ *efficient* at P if L is distributed as σN. It follows from a version of the convolution theorem that an estimator $\hat{\kappa}$ of $E[f(W)]$

is regular and efficient for $E[f(W)]$ at P if and only if

$$n^{1/2}(\hat{\kappa} - E[f(W)]) = n^{-1/2}\sum_{j=1}^{n} k(W_j) + o_p(1).$$

We refer to Bickel et al. (1998, Section 3.3, Theorem 2), for this characterization of efficient estimators. The empirical estimator $\hat{\kappa} = (1/n)\sum f(W_i)$ clearly obeys the characterization and is therefore regular and efficient.

We turn now to our model $W = (X, Y, Z)$ with Y and Z conditionally independent given X. This is a submodel of the above nonparametric model. In order to obtain a characterization of efficient estimators, we must determine the tangent space of this submodel (which is a subspace of $L_{2,0}(P)$), and the gradient of $E[f(W)]$ in this tangent space (which is the projection of the above gradient onto this tangent space).

The model can be parametrized in several ways, the simplest being the following. Since X and Z are conditionally independent given Y, we can write the distribution of $W = (X, Y, Z)$ as $P(dx, dy, dz) = S(dx, dy)R(dz|x)$. This means that P is conveniently parametrized by just two parameters, S and R. To prove local asymptotic normality, we introduce one-dimensional local models separately for S and R by setting

$$S_{na}(dx, dy) = S(dx, dy)(1 + n^{-1/2}a(x, y)),$$
$$R_{nb}(dz|x) = R(dz|x)(1 + n^{-1/2}b(x, z)).$$

Since S_{na} must be a probability distribution and R_{nb} a conditional distribution, we take a in the space A of bounded functions $a(x, y)$ with $Sa = 0$, and b in the space B of bounded functions $b(x, z)$ with $Rb(X) = 0$. The condition on b implies that $a(X, Y)$ and $b(X, Z)$ are orthogonal. Let P^n and P^n_{nab} denote the joint distribution of the observations $(W_1, \ldots, W_n)$ under $P(dx, dy, dz) = S(dx, dy)R(dz|x)$ and $P_{nab}(dx, dy, dz) = S_{na}(dx, dy)R_{nb}(dz|x)$, respectively. As above we have *local asymptotic normality* at P^n, now of the form

$$\log \frac{dP^n_{nab}}{dP^n} = n^{-1/2}\sum_{j=1}^{n} \big(a(X_j, Y_j) + b(X_j, Z_j)\big) - \frac{1}{2}\big(E[a^2(X, Y)] + E[b^2(X, Z)]\big) + o_p(n^{-1/2}).$$

Let $\bar{A}$ be the closure of A in $L_2(S)$ and $\bar{B}$ the closure of B in $L_2(T)$. The squared norm

$$\|(a, b)\|^2 = E[a^2(X, Y)] + E[b^2(X, Z)]$$

induces an inner product on the closure in $L_{2,0}(P)$ of the tangent space,

$$\bar{A} + \bar{B} = \{a(X, Y) + b(X, Z) : a \in \bar{A}, b \in \bar{B}\}.$$

Since a and b are orthogonal, the inner product decomposes as

$$((a, b), (a', b')) = E[a(X, Y)a'(X, Y)] + E[b(X, Z)b'(X, Z)].$$

A real-valued functional κ of (S, R) is *differentiable* at (S, R) with *gradient* $(g, h) \in$

$\bar{A} \times \bar{B}$ if

$$n^{1/2}(\kappa(S_{na}, R_{nb}) - \kappa(S, R)) \to ((g, h), (a, b)) \quad \text{for all} \quad (a, b) \in A \times B.$$

An estimator $\hat{\kappa}$ of κ is *regular* at (S, R) with *limit* L if L is a random variable such that

$$n^{1/2}(\hat{\kappa} - \kappa(S_{na}, R_{nb})) \Rightarrow L \quad \text{under } P_{nab} \quad \text{for all} \quad (a, b) \in A \times B.$$

As in the nonparametric case we obtain the following characterization. An estimator $\hat{\kappa}$ of κ is regular and efficient for κ at (S, R) if and only if

$$n^{1/2}(\hat{\kappa} - \kappa(S, R)) = n^{-1/2} \sum_{j=1}^{n} \big(g(X_j, Y_j) + h(X_j, Z_j)\big) + o_p(1). \tag{3.1}$$

We note that, by the Cramér–Wold device, the characterization (3.1) extends to d-dimensional κ with a multivariate version of regularity and a componentwise definition of differentiability. An efficient estimator for κ then has the asymptotic distribution $\Sigma^{1/2} N_d$ with N_d a d-dimensional standard normal random vector and $\Sigma = ((g, h), (g, h)^\top)$.

Here we are interested in the functional

$$\kappa(S, R) = E[f(W)] = \iint f(x, y, z) S(dx, dy) R(dz|x)$$

for bounded f. For $(a, b) \in A \times B$ we write

$$\begin{aligned} &n^{1/2}(\kappa(S_{na}, R_{nb}) - \kappa(S, R)) \\ &\qquad = n^{1/2} \iint f(x, y, z)\big(S_{na}(dx, dy) R_{nb}(dz|x) - S(dx, dy) R(dz|x)\big) \\ &\qquad \to \iint f(x, y, z)\big(a(x, y) + b(x, z)\big) S(dx, dy) R(dz|x). \end{aligned}$$

We rewrite the right-hand side as

$$\begin{aligned} &E[f(X, Y, Z)\big(a(X, Y) + b(X, Z)\big)] \\ &\qquad = E[Rf(X, Y) a(X, Y)] + E[Qf(X, Z) b(X, Z)] \\ &\qquad = E[\big(Rf(X, Y) - E[f(W)]\big) a(X, Y)] + E[\big(Qf(X, Z) - RQf(X)\big) b(X, Z)]. \end{aligned} \tag{3.2}$$

The last equality uses the fact that $Sa = 0$ and $Rb(X) = 0$. This shows that the gradient of $E[f(W)]$ at (S, R) is (g, h) with

$$g(X, Y) = Rf(X, Y) - E[f(W)], \qquad h(X, Z) = Qf(X, Z) - RQf(X).$$

Comparing with Theorem 2.1, we see that our plug-in estimator $\mathbb{T} = \mathbb{T}_f$ is regular and efficient for $E[f(W)]$ at (S, R) by characterization (3.1). It follows that $(\mathbb{T}_{f_1}, \ldots, \mathbb{T}_{f_d})$ is regular and efficient for $(E[f_1(W)], \ldots, E[f_d(W)])$ for each dimension d.

4. Simulations and discussion

In this article we propose a method of estimating distributions with conditionally independent components that actually uses the independence structure. The only competitor (which does not use the assumed independence and which therefore cannot be efficient) is the empirical estimator. In order to get a first idea about the practical performance of our estimator, we performed a small simulation study using `R`. Since our approach and the techniques are new, we have chosen the simplest boundary-adjusted kernel estimator, namely a linearly interpolated boundary kernel. This allowed us to keep the proofs short. For the simulations we chose the kernel more carefully. This makes a difference, especially when samples are very small, which is the situation here where we study samples of size $n = 30$. We revisit the examples from the Introduction, i.e., we consider estimating $P(Y < Z)$ and $H(y, z) = P(Y \le y, Z \le z)$. We take X to be uniformly distributed on $[-1, 1]$. The simulations are based on $10,000$ iterations.

In order to describe the estimators, and to avoid notational confusion, we will write ψ for the kernel k. Apart from that we will keep the notation from Section 2. For the simulations we chose the fourth-order kernel

$$\psi(x) = \frac{3 - x^2}{2}\,\phi(x), \quad x \in \mathbb{R},$$

where ϕ is the standard normal density. This kernel does not have a compact support, which we assume in Section 2 to prove our asymptotic results, but it works well in our case where the sample size is small: working with a higher order kernel with compact support $[-1, 1]$ would produce more frequently small or even negative values for the density estimators than a higher order kernel with infinite support. Now introduce

$$\Psi(u) = \int_{\infty}^{u} \psi(x)\,dx = \frac{1}{4}\Big(6\Phi(u) - 1 - \mathrm{sign(u)}\chi(u^2)\Big), \quad u \in \mathbb{R},$$

where Φ is the standard normal distribution function and χ is the chi-square distribution function with 3 degrees of freedom. Note that Ψ would be the distribution function associated with ψ if ψ were a proper (non-negative) probability density. Our boundary kernel h is

$$h(x) = \frac{2\pi}{\pi - 2}\Big(1 - \frac{2x}{\sqrt{2\pi}}\Big)\phi(x)\mathbf{1}[x \ge 0], \quad x \in \mathbb{R}.$$

Again, this kernel does not have a compact support, but it satisfies the integrability condition $\int xh(x)\,dx = 0$.

Let us begin with estimating $P(Y < Z)$. For our simulations the random variables Y given X and Z given X were generated (independently) from centered normal distributions with variances $1 + X^2$ and 1, respectively. For this choice $P(Y < Z) = 1/2$. For our kernel choice, the estimator from Example 2.2 computes to

$$\mathbb{T} = \int_{-1}^{1} \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \int \Psi\Big(\frac{Z_j - Y_i}{c} + u\Big)\psi(u)\,du K_b(x, X_i)K_b(x, X_j)\,\frac{dx}{\hat{m}(x)}.$$

In our simulations we used the simplified version

$$\mathbb{T}_1 = \int_{-1}^{1} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Psi\Big(\frac{Z_j - Y_i}{c}\Big) K_b(x, X_i) K_b(x, X_j) \, \frac{dx}{\hat{m}(x)}.$$

In the simulations the integral was approximated by a Riemann sum, using an equidistant partition of width 0.01 and the midpoint rule. Table 2 lists the simulated mean square errors (MSE) multiplied by the sample size $n = 30$ for several choices of bandwidth b and c. The simulated values should be close to the theoretical value $1/6 \simeq 0.167$. (See the Introduction for details.) Our simulated MSE for the empirical estimator was 0.253, which is close to the theoretical value $1/4 = 0.25$. For all choices of b and c, our estimator substantially improved on the empirical estimator.

Table 2. Simulated mean square errors multiplied by n for estimating $P(Y < X)$

	$c = 0.6$	$c = 0.75$	$c = 0.9$	$c = 1.05$	$c = 1.2$
$b = 0.30$	0.178	0.172	0.166	0.159	0.151
$b = 0.45$	0.176	0.172	0.166	0.159	0.152
$b = 0.60$	0.180	0.175	0.170	0.163	0.155
$b = 0.75$	0.184	0.179	0.174	0.167	0.159
$b = 0.90$	0.184	0.180	0.174	0.168	0.160

Our second example is the distribution function $H(y, z) = P(Y \le y, Z \le z)$. We consider exactly the situation from the Introduction, where the conditional distributions of Y and Z given X are normal with means X and X^2, respectively, and variances 1. In the Introduction we showed that, with regard to the asymptotic variance, our estimator clearly outperforms the empirical estimator. The theoretical relative variance reduction is given in Table 1 in the Introduction, for selected values of y and z. We are interested in comparing the theoretical values with their simulated analogs. For the simulations we again used the fourth-order kernel $\psi(x)$ introduced above. Our estimator (see Example 2.3) based on this kernel can be written as

$$\mathbb{T}(y, z) = \int_{-1}^{1} \hat{F}_S(y|x) \hat{G}_S(z|x) \frac{dx}{\hat{m}(x)},$$

where $\hat{F}_S$ and $\hat{G}_S$ are the smoothed conditional distribution functions defined by

$$\hat{F}_S(y|x) = \frac{1}{n} \sum_{j=1}^{n} \Psi\Big(\frac{y - Y_j}{c}\Big) K_b(x, X_j) \quad \text{and} \quad \hat{G}_S(z|x) = \frac{1}{n} \sum_{j=1}^{n} \Psi\Big(\frac{z - Z_j}{c}\Big) K_b(x, X_j),$$

for $x, y \in \mathbb{R}$. The simulated relative variance reduction $\tilde{\varrho}(y, z)$ is given in Table 3, for y and z in the set $\{-1.5, -1, -0.5, 0, 0.5, 1.0, 1.5\}$, and with bandwidths $b = 0.5$ and $c = 0.8$. We used the formula

$$\tilde{\varrho}(y, z) = \frac{\mathrm{MSE}_{\mathrm{emp}}(y, z) - \mathrm{MSE}_{\mathbb{T}}(y, z)}{\mathrm{MSE}_{\mathrm{emp}}(y, z)},$$

where $\mathrm{MSE}_{\mathbb{T}}(x,y)$ and $\mathrm{MSE}_{\mathrm{emp}}(x,y)$ denote the simulated MSE's of our estimator and the empirical estimator, respectively.

Table 3. Observed relative reduction amounts $\tilde{\varrho}(y,z)$

$y\backslash z$	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5
-1.5	0.82	0.76	0.69	0.61	0.51	0.41	0.31
-1.0	0.74	0.70	0.65	0.59	0.51	0.42	0.33
-0.5	0.66	0.63	0.61	0.57	0.52	0.44	0.37
0.0	0.59	0.57	0.56	0.55	0.49	0.44	0.37
0.5	0.50	0.49	0.51	0.51	0.45	0.40	0.34
1.0	0.42	0.43	0.46	0.45	0.41	0.36	0.30
1.5	0.35	0.37	0.40	0.41	0.38	0.33	0.28

Comparing the simulated values in Table 3 with the theoretical values in Table 1 we see that, perhaps surprisingly, our results are much better than the asymptotic theory suggests, especially when y and z are both non-negative. (This may be less pronounced with a data-driven choice of bandwidth.) We attribute the variance reduction to the fact that our estimator $\mathbb{T}$ uses *smoothed* conditional distribution functions $\hat{F}_S$ and $\hat{G}_S$, which typically behave better for small sample sizes than their unsmoothed versions $\hat{F}$ and $\hat{G}$ given by

$$\hat{F}(y|x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}[Y_j \le y] K_b(x, X_j) \quad \text{and} \quad \hat{G}(z|x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}[Z_j \le z] K_b(x, X_j).$$

This is indeed confirmed by further simulations, now using the unsmoothed version $\mathbb{T}_U$ of $\mathbb{T}$,

$$\mathbb{T}_U(y,z) = \int_{-1}^{1} \hat{F}(y|x)\hat{G}(z|x)\frac{dx}{\hat{m}(x)}$$

obtained by replacing $\hat{F}_S$ and $\hat{G}_S$ by $\hat{F}$ and $\hat{G}$. The simulated relative variance reductions are given in Table 4. We see that the values for this estimator are much closer to the asymptotic values of Table 1.

Table 4. Observed relative reduction amounts $\tilde{\varrho}(y,z)$ for the unsmoothed estimator

$y\backslash z$	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5
-1.5	0.77	0.72	0.63	0.49	0.35	0.19	0.07
-1.0	0.67	0.63	0.55	0.44	0.31	0.17	0.06
-0.5	0.55	0.52	0.46	0.39	0.28	0.16	0.07
0.0	0.43	0.41	0.36	0.32	0.24	0.16	0.07
0.5	0.31	0.29	0.27	0.24	0.18	0.13	0.05
1.0	0.21	0.20	0.18	0.15	0.13	0.09	0.03
1.5	0.12	0.12	0.11	0.09	0.07	0.05	0.01

5. Proof of Theorem 2.1

We need two auxiliary results on convergence rates for $\hat{m}$, $\hat{s}$ and $\hat{t}$.

Lemma 5.1: *Suppose m has support $[\alpha, \beta]$ and is positive and continuous on $[\alpha, \beta]$. Then $\hat{m}$ is a uniformly consistent estimator of m as $b \to 0$ and $nb^2 \to \infty$,*

$$\sup_{\alpha \le x \le \beta} |\hat{m}(x) - m(x)| = o_p(1).$$

If the restriction of m to $[\alpha, \beta]$ is differentiable and its derivative m' is Lipschitz, then we have the rate

$$\int_\alpha^\beta E[(\hat{m}(x) - m(x))^2]\, dx = O(b^4 + 1/(nb)).$$

Proof of Lemma 5.1: The uniform convergence follows from standard results, using the uniform continuity of m on $[\alpha, \beta]$; see Parzen (1962). To prove the second conclusion, we let

$$\bar{m}(x) = E[\hat{m}(x)] = E[K_b(x, X)].$$

Then we can write the left-hand side in the second assertion as the sum of the variance term $A = \int_\alpha^\beta E[(\hat{m}(x) - \bar{m}(x))^2]\, dx$ and the squared bias term $B = \int_\alpha^\beta (\bar{m}(x) - m(x))^2\, dx$. For x in $[\alpha, \beta]$ we have

$$nbE[(\hat{m}(x) - \bar{m}(x))^2] \le \|m\|_\infty \max\{\|k^2\|_1, \|h^2\|_1\}.$$

This follows from the fact that the variance of a convex combination of two random variables is bounded by the same convex combination of their variances, and from the inequality

$$nb\mathrm{Var}\Big(\frac{1}{nb}\sum_{j=1}^n \ell\Big(\frac{x - X_j}{b}\Big)\Big) \le \frac{1}{b} E\Big[\ell^2\Big(\frac{x - X}{b}\Big)\Big] = \int \ell^2(u) m(x - bu)\, du \le \|m\|_\infty \|l^2\|_1,$$

valid for every square-integrable function ℓ. This shows that $A = O(1/(nb))$.

Now we treat the squared bias B. For x in $[\alpha + b, \beta - b]$, we have the identity

$$\bar{m}(x) - m(x) = \int (m(x - bu) - m(x) - bum'(x))k(u)\, du$$

in view of the symmetry of k. By the assumption on m, there is a constant C such that

$$|m(y) - m(x) - (y - x)m'(x)| \le C|y - x|^2, \quad x, y \in [\alpha, \beta].$$

This yields

$$\sup_{x \in [\alpha + b, \beta - b]} |\bar{m}(x) - m(x)| = O(b^2).$$

Similarly, one verifies $|\bar{m}(x) - m(x)| = O(b^2)$ for $x = \alpha, \beta$. For $x \in (\alpha, \alpha + b)$ we derive

$$\begin{aligned} m(x) - \bar{m}(x) &= m(x) - \bar{m}(\alpha) - \frac{x-\alpha}{b}(\bar{m}(\alpha+b) - \bar{m}(\alpha)) \\ &= m(x) - m(\alpha) - (x-\alpha)m'(\alpha) + (m(\alpha) - \bar{m}(\alpha)) \\ &\quad - \frac{x-\alpha}{b}(m(\alpha+b) - m(\alpha) - bm'(\alpha)) \\ &\quad - \frac{x-\alpha}{b}(\bar{m}(\alpha+b) - m(\alpha+b) - \bar{m}(\alpha) + m(\alpha)). \end{aligned}$$

Thus, by the above, we obtain

$$\sup_{x \in [\alpha, \alpha+b]} |\bar{m}(x) - m(x)| = O(b^2).$$

The same result holds for the interval $[\beta - b, \beta]$. From this we conclude that $B = O(b^4)$.

Remark 3: Using the inequality

$$\Big(\int |\chi(y)| \, dy \Big)^2 \le \int \frac{1}{(1+|y|)^\eta} \, dy \int (1+|y|)^\eta \chi^2(y) \, dy,$$

valid for $\eta > 1$, and the identities

$$s(x+h, y) - s(x, y) - h\dot{s}_1(x, y) = h^2 \int \ddot{s}_{11}(x + \lambda h, y)(1 - \lambda) \, d\lambda$$

and

$$\dot{s}_1(x+h, y) - \dot{s}_1(x, y) = h \int \ddot{s}_{11}(x + \lambda h, y) \, d\lambda,$$

valid for x and $x+h$ in $[\alpha, \beta]$ and all real y, we see that the assumptions on s in Theorem 2.1 imply the smoothness assumptions on m in Lemma 5.1 with $m'(x) = \int \dot{s}_1(x, y) \, dy$. This is the reason why the smoothness assumptions of Lemma 5.1 are not mentioned in the assumptions of Theorem 2.1.

Lemma 5.2: *Suppose s is twice continuously differentiable on S with gradient $\dot{s}$ and Hessian $\ddot{s}$, and that s and $\|\ddot{s}\|^2$ belong to H_η for some $\eta \ge 0$. Then*

$$\|E[(\hat{s} - s)^2]\|_{H_\eta} = \sup_{\alpha \le x \le \beta} \int (1+|y|)^\eta E[(\hat{s}(x, y) - s(x, y))^2] \, dy = O(c^4 + b^4 + 1/(nbc)).$$

Proof of Lemma 5.2: Let $\bar{s}(x, y) = E[\hat{s}(x, y)]$. Also set

$$\begin{aligned} A(x) &= \int (1+|y|)^\eta E[(\hat{s}(x, y) - \bar{s}(x, y))^2] \, dy, \\ B(x) &= \int (1+|y|)^\eta (\bar{s}(x, y) - s(x, y))^2 \, dy. \end{aligned}$$

In what follows we need the following result. If ψ belongs to H_η and χ is an integrable function with support $[-1,1]$, then the function ψ_c defined by

$$\psi_c(x,y) = \int \psi(x, y - cv)\chi(v)\, dv$$

belongs to H_η in view of the inequality

$$\|\psi_c\|_{H_\eta} \le (1+c)^\eta \|\psi\|_{H_\eta} \|\chi\|_1. \tag{5.1}$$

The latter follows from the substitution $u = y - cv$, the inequality $1 + |x + y| \le (1 + |x|)(1 + |y|)$, valid for all x and y, and the fact that χ has support $[-1, 1]$.

Since $s(x, y) = 0$ for x outside the interval $[\alpha, \beta]$, we have for a square-integrable function ℓ,

$$\begin{aligned} nbc \int (1+|y|)^\eta \mathrm{Var}\Big(\frac{1}{nbc}\sum_{j=1}^n \ell\Big(\frac{x - X_j}{b}\Big) k\Big(\frac{y - Y_j}{c}\Big)\Big)\, dy \\ &\le \int (1+|y|)^\eta \frac{1}{bc} E\Big[\Big(\ell^2\Big(\frac{x-X}{b}\Big) k^2\Big(\frac{y-Y}{c}\Big)\Big]\, dy \\ &\le \iiint (1+|y|)^\eta s(x-bu, y-cv)\ell^2(u)k^2(v)\, du\, dv\, dy \\ &\le \|s\|_{H_\eta} \|\ell^2\|_1 (1+c)^\eta \|k^2\|_1. \end{aligned}$$

In the last step we have used (5.1) with $\chi = k^2$. From this and the definition of $\hat{s}$ we derive, for $x \in [\alpha, \beta]$,

$$A(x) \le \frac{(1+c)^\eta}{nbc} \|s\|_{H_\eta} \|k^2\|_1 \max\{\|h^2\|_1, \|k^2\|_1\}. \tag{5.2}$$

Now we treat $B(x)$. We recall that, for a twice continuously differentiable function on $[0, 1]$, we have

$$g(1) = g(0) + g'(0) + \int_0^1 g''(\lambda)(1-\lambda)\, d\lambda.$$

If the line segment $\{(x - \lambda bu, y - \lambda cv) : 0 \le \lambda \le 1\}$ belongs to S, then we have

$$s(x-bu, y-cv) = s(x,y) - (bu, cv)\dot{s}(x,y) + \int_0^1 (bu, cv)\ddot{s}(x - \lambda bu, y - \lambda cv)(bu, cv)^\top\, d\lambda.$$

Thus, for $x \in [\alpha + b, \beta - b]$, we have

$$\bar{s}(x,y) - s(x,y) = \iint \int_0^1 (bu, cv)\ddot{s}(x - \lambda bu, y - \lambda cv)(bu, cv)^\top (1-\lambda)\, d\lambda\, k(u)k(v)\, du\, dv.$$

Applying the Cauchy–Schwarz inequality, we have

$$(\bar{s}(x,y)-s(x,y))^2 \le 4(b^4+c^4)\|k\|_1^2 \int_0^1 \iint \|\ddot{s}(x-\lambda bu, y-\lambda cv)\|^2 |k(u)k(v)|\,du\,dv\,d\lambda.$$

Thus we obtain

$$\sup_{\alpha+b\le x\le\beta-b} B(x) \le 4(b^4+c^4)\|k\|_1^4(1+c)^\eta \|\|\ddot{s}\|^2\|_{H_\eta}.$$

A similar argument yields

$$\int (1+|y|)^\eta (\bar{s}(\alpha,y)-s(\alpha,y))^2\,dy \le 4(b^4+c^4)(1+c)^\eta \|\|\ddot{s}\|^2\|_{H_\eta}\|h\|_1^2\|k\|_1^2.$$

By these two inequalities,

$$\sup_{\alpha\le x\le\alpha+b}\int (1+|y|)^\eta (\bar{s}(\alpha,y)-s_b(\alpha,y))^2\,dy \le 4(b^4+c^4)(1+c)^\eta \|\|\ddot{s}\|^2\|_{H_\eta}\|k\|_1^2(\|k\|_1^2+\|h\|_1^2)$$

with

$$s_b(x,y) = s(\alpha,y) + \frac{x-\alpha}{b}(s(\alpha+b,y)-s(\alpha,y)).$$

In view of the identity

$$s(x,y) = s(\alpha,y) + (x-\alpha)\dot{s}_1(\alpha,y) + (x-\alpha)^2\int_0^1 \ddot{s}_{11}(\alpha+\lambda(x-\alpha),y)(1-\lambda)\,d\lambda,$$

we obtain, for x in $[\alpha,\alpha+b]$,

$$|s_b(x,y)-s(x,y)| \le b^2\int_0^1 (|\ddot{s}_{11}(x+\lambda(x-\alpha),y)| + |\ddot{s}_{11}(x+\lambda b,y)|)(1-\lambda)\,d\lambda.$$

From this we immediately obtain

$$\sup_{\alpha\le x\le\alpha+b}\int (s_b(x,y)-s(x,y))^2(1+|y|)^\eta\,dy \le 2b^4\|\ddot{s}_{11}^2\|_{H_\eta}.$$

Arguing similarly for the interval $[\beta-b,\beta]$, we finally arrive at the bound

$$\sup_{\alpha\le x\le\beta} B(x) = O(b^4+c^4). \tag{5.3}$$

The conclusion follows from (5.2) and (5.3).

Proof of Theorem 2.1: The properties of m imply that $\mu = \inf\{m(x) : \alpha \le x \le \beta\}$ is positive. From this and the first conclusion of Lemma 5.1 we obtain that $1/\hat{\mu} = O_p(1)$,

where $\hat{\mu} = \inf\{|\hat{m}(x)| : \alpha \le x \le \beta\}$. We set

$$D_1 = \iint Rf(x,y)(\hat{s}(x,y) - s(x,y))\,dx\,dy,$$
$$D_2 = \iint Qf(x,z)(\hat{t}(x,z) - t(x,z))\,dx\,dz,$$
$$D_3 = \int RQf(x)(\hat{m}(x) - m(x))\,dx$$

and show

$$\mathbb{T} = E[f(X,Y,Z)] + D_1 + D_2 - D_3 + o_p(n^{-1/2}), \tag{5.4}$$
$$D_1 = \frac{1}{n}\sum_{j=1}^{n} Rf(X_j,Y_j) - E[Rf(X,Y)] + o_p(n^{-1/2}), \tag{5.5}$$
$$D_2 = \frac{1}{n}\sum_{j=1}^{n} Qf(X_j,Z_j) - E[Qf(X,Z)] + o_p(n^{-1/2}), \tag{5.6}$$
$$D_3 = \frac{1}{n}\sum_{j=1}^{n} RQf(X_j) - E[RQf(X)] + o_p(n^{-1/2}). \tag{5.7}$$

To prove (5.4) we use the relations between s, t, m and q, r to write

$$\frac{\hat{s}(x,y)\hat{t}(x,z)}{\hat{m}(x)} = \frac{s(x,y)t(x,z)}{m(x)} + (\hat{s}(x,y) - s(x,y))r(z|x) + (\hat{t}(x,z) - t(x,z))q(y|x) - q(y|x)r(z|x)(\hat{m}(x) - m(x)) + \hat{R}(x,y,z)$$

with remainder term

$$\begin{aligned}\hat{R}(x,y,z) = {} & q(y|x)r(z|x)\Big(\frac{m^2(x)}{\hat{m}(x)} - \frac{m^2(x)}{m(x)} + \hat{m}(x) - m(x)\Big) \\ & + (\hat{s}(x,y) - s(x,y))(\hat{t}(x,z) - t(x,z))\frac{1}{\hat{m}(x)} \\ & + (\hat{s}(x,y) - s(x,y))\Big(\frac{1}{\hat{m}(x)} - \frac{1}{m(x)}\Big)m(x)r(z|x) \\ & + q(y|x)m(x)(\hat{t}(x,z) - t(x,z))\Big(\frac{1}{\hat{m}(x)} - \frac{1}{m(x)}\Big).\end{aligned}$$

To calculate $\mathbb{T}$, we integrate f against the four leading terms and the four remainder terms above. The four leading terms yield $E[f(X,Y,Z)] + D_1 + D_2 - D_3$. Thus (5.4) follows if we show that the integrals against the four remainder terms are of order $o_p(n^{-1/2})$. For this we use the following results which are consequences of Lemmas 5.1 and 5.2 and the

properties of b and c, see also Remark 2:

$$\int_\alpha^\beta \big(\hat m(x) - m(x)\big)^2\,dx = o_p(n^{-1/2}), \tag{5.8}$$

$$\int_\alpha^\beta \int (1+|y|)^\eta \big(\hat s(x,y) - s(x,y)\big)^2\,dy\,dx = o_p(n^{-1/2}), \tag{5.9}$$

$$\int_\alpha^\beta \int (1+|z|)^\eta \big(\hat t(x,z) - t(x,z)\big)^2\,dz\,dx = o_p(n^{-1/2}). \tag{5.10}$$

Since f is bounded by 1, the integral against the first remainder term is bounded by

$$\int \Big|\frac{m^2(x)}{\hat m(x)} - \frac{m^2(x)}{m(x)} + \hat m(x) - m(x)\Big|\,dx \le \frac{1}{\hat\mu}\int_\alpha^\beta \big(\hat m(x) - m(x)\big)^2\,dx = o_p(n^{-1/2}).$$

For two measurable functions γ and χ defined on $\mathbb{R}^2$, we have

$$\begin{aligned}
&\iiint |\gamma(x,y)\chi(x,z)|\,dx\,dy\,dz\\
&\qquad\le \iint \Big(\int \gamma^2(x,y)\,dx \int \chi^2(x,z)\,dx\Big)^{1/2} dy\,dz\\
&\qquad= \iint \Big(\int \gamma^2(x,y)\,dx \int \chi^2(x,z)\,dx\Big)^{1/2} \frac{(1+|y|)^{\eta/2}(1+|z|)^{\eta/2}}{(1+|y|)^{\eta/2}(1+|z|)^{\eta/2}}\,dy\,dz\\
&\qquad\le \int \frac{du}{(1+|u|)^\eta}\Big(\iint (1+|y|)^\eta\gamma^2(x,y)\,dx\,dy \iint (1+|z|)^\eta\chi^2(x,z)\,dx\,dz\Big)^{1/2}.
\end{aligned}$$

In view of this inequality and (5.9) and (5.10), the integral of f against the second remainder term is of order $o_p(n^{-1/2})$. For measurable functions γ on $\mathbb{R}^2$ and χ on $\mathbb{R}$, we have

$$\begin{aligned}
\iint |\gamma(x,y)\chi(x)|\,dx\,dy &\le \int \Big(\int \gamma^2(x,y)\,dx \int \chi^2(x)\,dx\Big)^{1/2} dy\\
&\le \Big(\int \frac{dx}{(1+|x|)^\eta} \iint (1+|y|)^\eta\gamma^2(x,y)\,dx\,dy \int \chi^2(x)\,dx\Big)^{1/2}.
\end{aligned}$$

In view of this inequality, (5.8)–(5.10), and $1/(\hat\mu\mu) = O_p(1)$, the integrals against the last two remainder terms are of order $o_p(n^{-1/2})$. This completes the proof of (5.4).

To prove (5.5), we set $g = Rf$ and

$$g_n(x,y) = \iint g(u, y+cv)K_b(x,u)k(v)\,du\,dv.$$

With $\bar{s}(x,y) = E[\hat{s}(x,y)]$ we have

$$\iint g(x,y)\big(\hat{s}(x,y) - \bar{s}(x,y)\big)\,dx\,dy = \frac{1}{n}\sum_{j=1}^{n}\big(g_n(X_j,Y_j) - E[g_n(X,Y)]\big).$$

It follows from (2.1) that

$$E[\big(g_n(X,Y) - g(X,Y)\big)^2] \to 0.$$

Indeed, since $|g| \le 1$, we have $|g_n| \le C = \|k\|_1(\|h\|_1 + \|k\|_1)$ and can bound the left-hand side by $(1+C)^2(1 - E[I_b]) + B_n$, where

$$\begin{aligned} B_n &= E\Big[I_b\Big(\iint \big(g(X+bu, Y+cv) - g(X,Y)\big)k(u)k(v)\,du\,dv\Big)^2\Big] \\ &\le \|k\|_1^2 E\Big[I_b \iint \big(g(X+bu,Y+cv) - g(X,Y)\big)^2 |k(u)k(v)|\,du\,dv\Big] \end{aligned}$$

converges to zero by (2.1). The above shows that

$$\Delta = \iint g(x,y)\big(\hat{s}(x,y) - \bar{s}(x,y)\big)\,dx\,dy - \frac{1}{n}\sum_{j=1}^{n}\big(g(X_j,Y_j) - E[g(X,Y)]\big) = o_p(n^{-1/2}).$$

Here we used $nE[\Delta^2] \le E[(g_n(X,Y) - g(X,Y))^2] \to 0$.

To obtain (5.5) it suffices to verify $E[g_n(X,Y)] = E[g(X,Y)] + o(n^{-1/2})$. In view of (2.2) this follows if we show

$$\begin{aligned} \Gamma &= E[(1-I_b)(g_n(X,Y) - g(X,Y))] \\ &= \iint \mathbf{1}[x \notin [\alpha+2b, \beta-2b]]g(x,y)\big(\bar{s}(x,y) - s(x,y)\big)\,dx\,dy = o(n^{-1/2}). \end{aligned}$$

An application of the Cauchy–Schwarz inequality yields

$$\int |\bar{s}(x,y) - s(x,y)|\,dy \le \Big(\int \frac{dy}{(1+|y|)^\eta} B(x)\Big)^{1/2},$$

where $B(x) = \int (1+|y|)^\eta(\bar{s}(x,y) - s(x,y))^2\,dy$. From this, $|g| \le 1$ and (5.3) we obtain $\Gamma = O(b^3 + bc^2)$. This completes the proof of (5.5). The proofs of (5.6) and (5.7) are similar and will be omitted.

Acknowledgements

Ursula U. Müller was supported by NSF Grant DMS 0907014. Anton Schick was supported by NSF Grant DMS 0906551. The authors thank the referees and an Associate Editor for a number of suggestions that improved the manuscript.

References

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J.A. (1998), *Efficient and Adaptive Estimation for Semiparametric Models*, Springer, New York.

Bouezmarni, T., and Rombouts, J.V.K. (2010), 'Nonparametric Density Estimation for Multivariate Bounded Data', *Journal of Statistical Planning and Inference*, 140, 139–152.

Causeur, D., and Dhorne, T. (2003), 'Linear Regression Models Under Conditional Independence Restrictions', *Scandinavian Journal of Statistics*, 30, 637–650.

Delgado, M., Domínguez, M., and Lavergne, P. (2006), 'Consistent Tests of Conditional Moment Restrictions', *Annals of Economics and Statistics*, 81, 33–67.

Delgado, M.A., and González Manteiga, W. (2001), 'Significance Testing in Nonparametric Regression Based on the Bootstrap', *The Annals of Statistics*, 29, 1469–1507.

Efromovich, S. (1999), *Nonparametric Curve Estimation*, Springer Series in Statistics, Springer-Verlag, New York.

Efromovich, S. (2005), 'Oracle Inequality for Conditional Density Estimation and an Actuarial Example', *Annals of the Institute of Statistical Mathematics*, 62, 249–275.

Efromovich, S. (2010), 'Dimension Reduction and Adaptation in Conditional Density Estimation', *Journal of the American Statistical Society*, 105, 761–774.

Fernandes, M., and Flôres Jr., E.G. (1999), 'Tests for Conditional Independence, Markovian Dynamics, and Noncausality', Discussion Paper, European University Institute.

Hájek, J. (1970), 'A Characterization of Limiting Distributions of Regular Estimates', *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14, 323–330.

Huang, T.M. (2010), 'Testing Conditional Independence Using Maximal Nonlinear Conditional Correlation', *The Annals of Statistics*, 38, 2047–2091.

Koshevnik, Yu.A., and Levit, B.Ya, (1976), 'On a Non-parametric Analogue of the Information Matrix', *Theory of Probability and its Applications*, 21, 759–774.

Lauritzen, S.L. (1996), *Graphical Models*, Oxford Statistical Science Series 17, Oxford University Press, New York.

Le Cam, L. (1956), *On the Asymptotic Theory of Estimation and Testing Hypotheses*, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, pp. 129–156.

Le Cam, L. (1966), *Likelihood Functions for Large Numbers of Independent Observations*, in *Research Papers in Statistics, Festschrift J. Neyman*, Wiley, London, pp. 167–187.

Le Cam, L. (1969), *Théorie Asymptotique de la Décision Statistique*, Séminaire de Mathématiques Supérieures 33, Les Presses de l'Université de Montréal, Montreal.

Le Cam, L. (1972), *Limits of Experiments*, in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, pp. 245–261.

Linton, O., and Gozalo, P. (1999), 'Conditional Independence Restrictions: Testing and Estimation', *http://econ.lse.ac.uk/staff/olinton/research/Misc/cindtr.pdf*.

Masri, R., and Redner, R.A. (2005), 'Convergence Rates for Uniform B-spline Density Estimators on Bounded and Semi-infinite Domains', *Journal of Nonparametric Statistics*, 17, 555–582.

Marshall, J.C., and Hazelton, M.L. (2010), 'Boundary Kernels for Adaptive Density Estimators on Regions with Irregular Boundaries', *Journal of Multivariate Analysis*, 101, 949–963.

Müller, H.-G. (1993), 'On the Boundary Kernel Method for Nonparametric Curve Estimation near Endpoints', *Scandinavian Journal of Statistics*, 20, 313–328.

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

Müller, U.U., Schick, A., and Wefelmeyer, W. (2005), 'Weighted Residual-based Density Estimators for Nonlinear Autoregressive Models', *Statistica Sinica*, 15, 177–195.

Parzen, E. (1962), 'On Estimation of a Probability Density Function and Mode', *The Annals of Mathematical Statistics*, 33, 1065–1076.

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*, Cambridge University Press, Cambridge.

Schick, A., and Wefelmeyer, W. (2004), 'Functional Convergence and Optimality of Plug-in Estimators for Stationary Densities of Moving Average Processes', *Bernoulli*, 10, 889–917.

Schick, A., and Wefelmeyer, W. (2007), 'Root-n Consistent Density Estimators of Convolutions in Weighted L_1-Norms', *Journal of Statistical Planning and Inference*, 137, 1765–1774.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Stein, C. (1956), *Efficient Nonparametric Testing and Estimation*, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, pp. 187–195.

Su, L., and White, H. (2007), 'A Consistent Characteristic Function-based Test for Conditional Independence', *Journal of Econometrics*, 141, 807–834.

Su, L., and White, H. (2008), 'A Nonparametric Hellinger Metric Test for Conditional Independence', *Econometric Theory*, 24, 829–864.