Hall, Peter H.; Horowitz, Joel L.

Working Paper

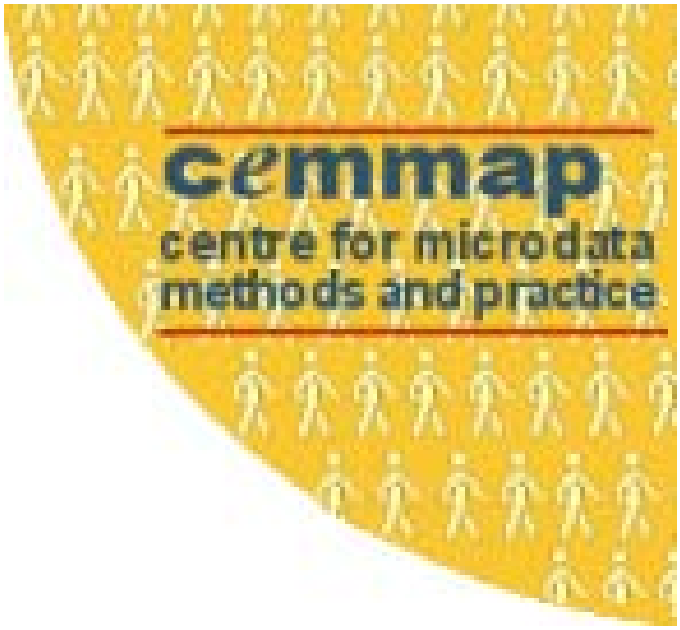
Nonparametric methods for inference in the presence of instrumental values

cemmap working paper, No. CWP02/03

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Hall, Peter H.; Horowitz, Joel L. (2003) : Nonparametric methods for inference in the presence of instrumental values, cemmap working paper, No. CWP02/03, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2003.0203

This Version is available at:
https://hdl.handle.net/10419/79372



NONPARAMETRIC METHODS FOR INFERENCE IN THE PRESENCE OF INSTRUMENTAL VARIABLES

Peter Hall
Joel L. Horowitz

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP02/03

NONPARAMETRIC METHODS FOR INFERENCE IN THE PRESENCE OF INSTRUMENTAL VARIABLES

Peter Hall[1] Joel L. Horowitz[2]

ABSTRACT. We suggest two nonparametric approaches, based on kernel methods and orthogonal series, respectively, to estimating regression functions in the presence of instrumental variables. For the first time in this class of problems we derive optimal convergence rates, and show that they are attained by particular estimators. In the presence of instrumental variables the relation that identifies the regression function also defines an ill-posed inverse problem, the "difficulty" of which depends on eigenvalues of a certain integral operator which is determined by the joint density of endogenous and instrumental variables. We delineate the role played by problem difficulty in determining both the optimal convergence rate and the appropriate choice of smoothing parameter.

KEYWORDS. Bandwidth, convergence rate, eigenvalue, endogenous variable, exogenous variable, kernel method, linear operator, nonparametric regression, smoothing, optimality.

SHORT TITLE. Instrumental variables.

AMS SUBJECT CLASSIFICATION: Primary, 62G08; Secondary, 62G20.

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

[2] Department of Economics, Anderson Hall, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208-2600, USA. The research of Joel L. Horowitz was supported in part by NSF Grant SES 9910925.

1. INTRODUCTION

Data (X_i, Y_i) are observed, the pairs being generated by the model

$$Y_i = g(X_i) + U_i \,, \tag{1.1}$$

where g is a function which we wish to estimate and the U_i's denote disturbances. The U_i's are correlated with the explanatory variables X_i, and in particular, $E(U_i \mid X_i)$ does not vanish. For example, this may occur if a third variable causes both X_i and Y_i, but is not included in the model.

This circumstance arises frequently in economics. To illustrate, suppose that Y_i denotes the hourly wage of individual i, and that X_i includes the individual's level of education, among other variables. The "error" U_i would generally include personal characteristics, such as "ability," which influence the individual's wage but are not observed by the analyst. If high-ability individuals tend to choose high levels of education, then education is correlated with ability, thereby causing U_i to be correlated with at least some components of X_i.

Suppose, however, that for each i we have available another observed data value, W_i say (an instrumental variable), for which

$$E(U_i \mid W_i) = 0 \tag{1.2}$$

and there is a "sufficiently strong" relationship between X_i and W_i. Then there is an opportunity for estimating g from the data (X_i, W_i, Y_i).

The formal definition of "sufficiently strong" will depend on the nature of the problem. In a parametric setting, for example where $g(X_i) = X_i\beta$, X_i is an $m \times k$ matrix and β is a $k \times 1$ vector, "sufficiently strong" means simply that the matrix of correlations between X and W is of full rank; this is sometimes expressed as "X and W are fully correlated." In a nonparametric setting the definition of "sufficiently strong" is given by, for example, condition (2.1) below.

Estimation of g is difficult because, as explained in section 2, the relation that identifies g is a Fredholm equation of the first kind, which leads to an ill-posed inverse problem (O'Sullivan, 1986; Kress, 1999). We use a ridge-type regularisation method to achieve boundedness of the relevant inverse integral operator, and develop both kernel and series estimators of g. The resulting estimators have optimal L_2 rates of convergence.

Related research on this problem is mostly very recent. Blundell and Powell (2003) and Florens (2003) discussed the relationship between (1.1) and other "structural" models in econometrics. Newey, Powell and Vella (1999) investigated estimation and inference with a triangular-array version of (1.1). In that set-up, equations relate X_i and W_i, and the disturbances of these equations are connected to U_i. Newey and Powell (2002) proposed a series estimator for g in (1.1), and gave sufficient conditions for its consistency but did not obtain a rate of convergence. Darolles, Florens and Renault (2002) developed a kernel estimator for a special case of (1.1) and obtained its rate of convergence. This rate is slower than that obtained here, and, therefore, suboptimal; see section 4.3 below.

Further, related work on inverse problems includes that of Wahba (1973), Tikhonov and Arsenin (1977), Groetsch (1984), Nashed and Wahba (1984) and Van Rooij and Ruymgaart 1999).

We shall give a relatively detailed treatment, together with proofs, of results in the case where the instrumental variable is univariate. This setting is arguably of greatest interest to statisticians. Extensions to multivariate cases will be outlined.

2. Model and estimators in bivariate case

2.1. Model. Let (U_i, W_i, X_i, Y_i), for $i \geq 1$, be independent and identically distributed 4-vectors, and assume they follow a model satisfying (1.1) and (1.2). We shall suppose that (W_i, X_i, Y_i), for $1 \leq i \leq n$, are observed, and that the distribution of (X_i, W_i) is confined to the unit square; if it is not then a monotone transformation will achieve this end.

Denote by f_X, f_W and f_{XW} the marginal densities of X and W, and the joint density of X and W, respectively, and define the linear operator T, on the space of square-integrable functions on $[0,1]^2$, by

$$(T\psi)(z) = \int t(x,z)\,\psi(x)\,dx\,,$$

where

$$t(x,z) = \int f_{XW}(x,w) f_{XW}(z,w)\,dw\,.$$

The following assumption characterises the strength of association we require between X and W:

$$T \quad \text{is invertible.} \tag{2.1}$$

To appreciate the nature of (2.1), observe that if X and W are independent then T maps each function ψ to a constant multiple of f_X, and so (2.1) fails. However, if (2.1) holds, then since it may be proved from (1.1) that

$$E_W\{E(Y \mid W)\, f_{XW}(z, W)\} = (Tg)(z)\,, \tag{2.2}$$

g may be recovered by inversion of T:

$$g(z) = E_W\big\{E(Y \mid W)\, \big(T^{-1} f_{XW}\big)(z, W)\big\}\,. \tag{2.3}$$

This property suggests an estimator, which we shall develop in section 2.2.

Observe that (2.2) is a Fredholm equation of the first kind, and generates an ill-posed inverse problem if, as is usually the case, zero is a limit point of the eigenvalues of T. In that case, T^{-1} is not a bounded, continuous operator. For the purpose of estimation, we shall deal with this problem in section 2.2 by replacing T^{-1} by $(T + a_n)^{-1}$, where a_n is a positive ridge parameter converging to zero as $n \to \infty$.

2.2. Generalised kernel estimator. Let $K_h(\cdot, \cdot)$ denote a generalised kernel function, with the property:

$$\text{for all } t \in [0,1]\,, \quad h^{-(j+1)} \int_{t-1}^{t} u^j\, K_h(u,t)\, du = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } j = 1\,. \end{cases} \tag{2.4}$$

Here, $h > 0$ denotes a bandwidth, and the kernel is considered in generalised form only to overcome edge effects. In particular, if h is small and t is not close to either 0 or 1 then we may take $K_h(u,t) = K(u/h)$, where K is a bounded, compactly supported, symmetric probability density. If t is close to 1 then we may take $K_h(u,t) = L(u/h)$, where L is a bounded, compactly supported function satisfying

$$\int_0^{\infty} u^j\, L(u)\, du = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } j = 1\,. \end{cases}$$

And if t is close to 0 then we may take $K_h(u,t) = L(-u/h)$. There are, of course, other ways of overcoming the edge-effect problem, but the "boundary kernel" approach above is also appropriate.

We require two estimators of f_{XW}, the second a leave-one-out estimator:

$$\hat{f}_{XW}(x,w) = \frac{1}{nh^2} \sum_{i=1}^{n} K_h(x - X_i, x)\, K_h(w - W_i, w)\,,$$

$$\hat{f}^{(-i)}_{XW}(x,w) = \frac{1}{(n-1)h^2} \sum_{1 \le j \le n\,:\, j \ne i} K_h(x - X_j, x)\, K_h(w - W_j, w)\,.$$

We use $\hat{f}_{XW}$ to construct the following estimators of $t(x,z)$ and the transformation T:

$$\hat{t}(x,z) = \int \hat{f}_{XW}(x,w)\,\hat{f}_{XW}(z,w)\,dw\,, \quad (\widehat{T}\psi)(z) = \int \hat{t}(x,z)\,\psi(x)\,dx\,.$$

Let $a_n > 0$; we shall used it as a ridge parameter when inverting $\widehat{T}$, defining $\widehat{T}^+ = (\widehat{T} + a_n)^{-1}$. Reflecting (2.3), our estimator of g is:

$$\hat{g}(x) = \frac{1}{n} \sum_{i=1}^{n} \bigl(\widehat{T}^+ \hat{f}_{XW}^{(-i)}\bigr)(z, W_i)\,Y_i\,.$$

2.3. Orthogonal series estimator. This technique is based on empirically transforming the marginal distributions of W and X to uniform, and exploiting the relatively simple character of the problem in that case. To appreciate this point, assume for the time being that both marginals are in fact uniform on $[0,1]$, and let $\chi_1, \chi_2, \ldots$ denote an orthonormal basis for $L_2[0,1]$. In practice one would usually take $\{\chi_j\}$ to be the cosine sequence, although there are many other options.

Let $f_{XW}(x,w) = \sum_j \sum_k q_{jk}\,\chi_j(x)\,\chi_k(w)$ denote the generalised Fourier expansion of f_{XW}, and put $Q = (q_{jk})$, $p_j = E\{Y\,\chi_j(W)\}$, $\gamma_j = E\{g(X)\,\chi_j(X)\}$, $p = (p_j)$ and $\gamma = (\gamma_j)$, the latter two quantities being column vectors. By (1.1) and (1.2), $QQ^{\mathrm{T}}\gamma = Qp$, and therefore, $\gamma = (QQ^{\mathrm{T}})^{-1}Qp$. (This is really another way of writing (2.3); observe that the operator T takes g to a function of which the jth Fourier coefficient is $(QQ^{\mathrm{T}}\gamma)_j$.) Hence, the problem of estimating the Fourier coefficients γ_j of g reduces to one of estimating p_j and q_{jk}.

Next we describe how to solve the latter problem in general cases, where marginal distributions are not uniform. First transform the marginals, by computing $\widehat{W}_i = \widehat{F}_W(W_i)$ and $\widehat{X}_i = \widehat{F}_W(X_i)$, where $\widehat{F}_W$ and $\widehat{F}_X$ denote the empirical distribution functions of the data $W_1, \ldots, W_n$ and $X_1, \ldots, X_n$, respectively. Put $\hat{q}_{jk} = n^{-1} \sum_i \chi_j(\widehat{W}_i)\,\chi_k(\widehat{X}_i)$ and $\hat{p}_j = n^{-1} \sum_i \chi_j(\widehat{W}_i)\,Y_i$. Let $\widehat{Q}$ be the $m \times m$ matrix that has $\hat{q}_{jk}$ in position (j,k), and set

$$\widehat{\gamma} = (\widehat{\gamma}_j) = \bigl(\widehat{Q}\widehat{Q}^{\mathrm{T}} + a_n I_m\bigr)^{-1}\,\widehat{Q}\,\hat{p}\,,$$

where a_n denotes a ridge parameter and I_m is the $m \times m$ identity. Our estimator of g is

$$\bar{g}(x) = \sum_{j=1}^{m} \widehat{\gamma}_j\,\chi_j(x)\,.$$

In this estimator the number of terms, m, in the approximating Fourier series is the main smoothing parameter. It is relatively awkward to derive theory for the orthogonal series method, owing to the fact that the transformed data $\widehat{W}_i$ and $\widehat{X}_i$ are not independent, and to the difficulty of dealing theoretically with the large random matrix $\widehat{Q}$. Nevertheless, we shall show in section 4 that, under restrictions, the orthogonal series technique has optimal performance.

3. Model and estimators in multivariate case. In the model at (1.1) the explanatory variable X is endogenous, i.e. determined within the model. When the model is multivariate there is an opportunity for dividing the explanatory variable, which is now a vector, into two parts, one endogenous and the other determined outside the model, or exogenous.

We take (Y, X, Z, W, U) to be a vector, where Y and U are scalars, X and W are supported on $[0,1]^p$, and Z is supported on $[0,1]^q$. Generalising (1.1) and (1.2), the model is

$$Y_i = g(X_i, Z_i) + U_i\,, \quad E(U_i \mid Z_i, W_i) = 0\,,$$

where $(Y_i, X_i, Z_i, W_i, U_i)$, for $i \geq 1$, are independent and identically distributed as (Y, X, Z, W, U). Thus, X and Z are endogenous and exogenous explanatory variables, respectively. Data (Y_i, X_i, Z_i, W_i), for $1 \leq i \leq n$, are observed.

Let f_{XZW} denote the density of (X, Z, W), write f_Z for the density of Z, and for each $x_1, x_2 \in [0,1]^p$, put

$$t_z(x_1, x_2) = \int f_{XZW}(x_1, z, w)\, f_{XZW}(x_2, z, w)\, dw\,,$$

the analogue of $t(x_1, x_2)$ in section 2. Define the operator T_z on $L_2[0,1]^p$ by

$$(T_z\psi)(x) = \int t_z(\xi, x)\, \psi(\xi)\, d\xi\,.$$

Analogously to (2.3) it may be proved that, for each z for which T_z^{-1} exists,

$$g(x, z) = f_Z(z)\, E_{W|Z}\Big\{E(Y \mid Z = z, W)\, \big(T_z^{-1} f_{XZW}\big)(x, z, W) \;\Big|\; Z = z\Big\}\,,$$

where $E_{W|Z}$ denotes the expectation operator with respect to the distribution of W conditional on Z. In this formulation, $(T_z^{-1} f_{XZW})(x, z, W)$ denotes the result of applying T_z^{-1} to the function $f_{XZW}(\,\cdot\,, z, W)$, and evaluating the resulting function at x.

To construct an estimator of $g(x,z)$, given $h>0$ and p-vectors $x=(x^{(1)},\ldots,x^{(p)})$ and $\xi=(\xi^{(1)},\ldots,\xi^{(p)})$, let $K_{p,h}(x,\xi)=\prod_{1\le j\le p}K_h(x^{(j)},\xi^{(j)})$, put $K_{q,h}(z,\zeta)$ analogously for q-vectors z and ζ, let $h_x,h_z>0$, and define

$$\hat f_{XZW}(x,z,w)=\frac{1}{nh_x^{2p}h_z^q}\sum_{i=1}^n K_{p,h_x}(x-X_i,x)\,K_{q,h_z}(z-Z_i,z)\,K_{p,h_x}(w-W_i,w)\,,$$

$$\hat f^{(-i)}_{XZW}(x,z,w)=\frac{1}{(n-1)h_x^{2p}h_z^q}\sum_{1\le j\le n\,:\,j\ne i}K_{p,h_x}(x-X_j,x)\,K_{q,h_z}(z-Z_j,z)\times K_{p,h_x}(w-W_j,w)\,,$$

$$\hat t_z(x_1,x_2)=\int\hat f_{XZW}(x_1,z,w)\,\hat f_{XZW}(x_2,z,w)\,dw$$

and

$$(\widehat T_z\psi)(x,z,w)=\int\hat t_z(\xi,x)\,\psi(\xi,z,w)\,d\xi\,,$$

where ψ is a function from $\mathbb{R}^{2p+q}$ to the real line. Then the estimator of $g(x,z)$ is

$$\hat g(x,z)=\frac1n\sum_{i=1}^n\big(\widehat T_z^+\hat f^{(-i)}_{XZW}\big)(x,z,W_i)\,Y_i\,K_{q,h_z}(z-Z_i,z)\,.$$

4. Theoretical properties

4.1. Kernel method for bivariate case. The invertibility of T is central to our ability to successfully resolve g from data, and so it comes as no surprise to find that rates of convergence of estimators of g hinge on the rate at which the eigenvalues of T, say $\lambda_1\ge\lambda_2\ge\ldots>0$, converge to 0. Therefore, our regularity conditions will be framed in terms of an eigen-expansion representation of T. To this end, let ϕ_j denote an eigenfunction of T with eigenvalue λ_j, normalised so that $\phi_1,\phi_2,\ldots$ is an orthonormal basis for the space of square-integrable functions on the interval $[0,1]$. Then we may write:

$$t(x,z)=\sum_{j=1}^\infty\lambda_j\,\phi_j(x)\,\phi_j(z)\,,\quad f_{XW}(x,z)=\sum_{j=1}^\infty\sum_{k=1}^\infty d_{jk}\,\phi_j(x)\,\phi_k(z)\,,\quad g(x)=\sum_{j=1}^\infty b_j\,\phi_j(x)\,,\tag{4.1}$$

where d_{jk} and b_j denote generalised Fourier coefficients of f_{XW} and g, respectively.

Next we state regularity conditions. Assumption A.1 is equivalent to the intersection of (1.1) and (1.2); A.3 gives smoothness conditions, expressed through

the eigen-expansion of T; and A.4 describes the sizes of tuning parameters. The invertibility condition, (2.1), follows from the fact that each $\lambda_j > 0$, implied by A.3.

Let $C > 0$ be an arbitrary large, but fixed, constant, let $\alpha, \beta > 0$, and denote by $\mathcal{G} = \mathcal{G}(C, \alpha, \beta)$ the class of distributions G of (X, W, Y) that satisfy the following conditions.

A.1. The data (X_i, W_i, Y_i) are independent and identically distributed as (X, W, Y), where (X, W) is supported on $[0,1]^2$ and $E\{Y - g(X) \mid W = w\} \equiv 0$.

A.2. The distribution of (X, W) has a density, f_{XW}, with two derivatives (when viewed as a function restricted to $[0,1]^2$) bounded uniformly, in absolute value, by C; and the functions $E(Y^2 \mid W = w)$ and $E(Y^2 \mid X = x, W = w)$ are bounded uniformly by C.

A.3. The constants α, β satisfy $\alpha > 1$ and $\alpha \le \frac{1}{3}\beta + \frac{1}{2}$, and moreover, $|b_j| \le C\, j^{-\beta}$, $j^{-\alpha} \le C\, \lambda_j$ and $\sum_{k \ge 1} |d_{jk}| \le C\, j^{-\alpha/2}$ for all $j \ge 1$.

A.4. The parameters a_n and h satisfy $a_n \asymp n^{-\alpha/(2\beta+\alpha)}$ and $h \asymp n^{-2\beta/\{3(2\beta+\alpha)\}}$ as $n \to \infty$, where $c_n \asymp d_n$, for positive constants c_n and d_n, means that c_n/d_n is bounded away from zero and infinity.

A.5. The function $K_h(\,\cdot\,,\,\cdot\,)$ satisfies (2.4); for each $t \in [0,1]$, $K_h(h\,\cdot\,, t)$ is supported on $[(t-1)/h, t/h] \cap \mathcal{K}$, where $\mathcal{K}$ is a compact interval not depending on t; and

$$\sup_{h>0,\, t\in[0,1],\, u\in\mathcal{K}} |K_h(hu, t)| < \infty\,.$$

Theorem 4.1. *As $n \to \infty$,*

$$\sup_{G\in\mathcal{G}} \int_0^1 E_G\{\hat{g}(t) - g(t)\}^2\, dt = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big)\,.$$

More generally, it may be proved that if a particular distribution of (X, W, Y) satisfies A.1, and if $E(Y^2) < \infty$ and the density f_{XW} is continuous on $[0,1]$, then a_n and h can be chosen so that $\int E_G(\hat{g} - g)^2 \to 0$ as $n \to \infty$. Similar results, guaranteeing consistent estimation but without a convergence rate, may be derived in the settings of sections 4.2 and 4.3.

4.2. Orthogonal series method for bivariate case. We shall simplify theory by assuming the Fourier coefficients q_{jk} satisfy a strong diagonality condition. Under this assumption it is sufficient to work with a strongly diagonal form of $\widehat{Q}$, where

we redefine $\hat{q}_{jk} = 0$ if $|j-k| \geq N$ (where N is permitted to increase slowly with n), and leave $\hat{q}_{jk}$ unchanged otherwise. With this alteration to $\hat{q}_{jk}$, let $\widehat{Q} = (\hat{q}_{jk})$ be the indicated $m \times m$ matrix.

Recall from section 2.3 that $\chi_1, \chi_2, \ldots$ is an orthonormal basis for $L_2[0,1]$. Let F_W and F_X denote the marginal distribution functions of W and X, put $\widetilde{W} = F_W(W)$ and $\widetilde{X} = F_X(X)$, and let $f_{\widetilde{W}\widetilde{X}}$ denote the joint density of $(\widetilde{W}, \widetilde{X})$. Write $f_{\widetilde{W}\widetilde{X}}(w,x) = \sum_j \sum_k q_{jk} \chi_j(x)\, \chi_k(w)$ and $g(x) = \sum_j \gamma_j\, \chi_j(x)$ for the generalised Fourier transforms of these functions. Recall that we require the transformation represented by QQ^{T} to be invertible, so we may define $Q^{-1} = (q_{jk}^{(-1)})$ to be a generalised inverse of Q.

Given constants $\alpha \geq 2$, $\beta \geq \frac{1}{2}$ and $C_1, C_2 > 0$, let $\mathcal{H} = \mathcal{H}(C_1, C_2, \alpha, \beta)$ denote the class of distributions G of $(\widetilde{W}, \widetilde{X}, Y)$ for which:

$$E\{Y - g(\widetilde{X}) \mid \widetilde{W} = w\} \equiv 0\,, \qquad |q_{jk}| \leq C_1 \{\max(j,k)\}^{-\alpha/2} \exp(-C_2\,|j-k|)\,,$$
$$|q_{jk}^{(-1)}| \leq C_1 \{\max(j,k)\}^{\alpha/2} \exp(-C_2\,|j-k|)\,, \quad |p_j| \leq C_1\, j^{-\beta}, \quad E\big(Y^4\big) < C_1\,,$$

where the bounds are assumed to hold uniformly in $1 \leq j, k < \infty$.

Theorem 4.2. *Let $\{\chi_j\}$ denote the orthonormalised version of the cosine series on $[0,1]$. Take $\alpha \geq 2$ and $\beta \geq \frac{1}{2}$, and assume $a_n \asymp m^{-\alpha}$, $m \asymp n^{1/(2\beta+\alpha)}$, $N/\log n \to \infty$ and $N = O(n^{\epsilon})$ for all $\epsilon > 0$. Then, as $n \to \infty$,*

$$\sup_{G \in \mathcal{H}} \int_0^1 E_G(\bar{g} - g)^2 = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big)\,.$$

4.3. Kernel method for multivariate case. For each $z \in [0,1]^q$, let $\{\phi_{z1}, \phi_{z2}, \ldots\}$ denote the orthonormalised sequence of eigenvectors, and $\lambda_{z1} \geq \lambda_{z2} \geq \ldots > 0$ the respective eigenvalues, of the operator T_z. Assume that $\{\phi_{zj}\}$ forms an orthonormal basis of $L_2[0,1]^p$. Analogously to (4.1),

$$t_z(x_1, x_2) = \sum_{j=1}^{\infty} \lambda_{zj}\, \phi_{zj}(x_1)\, \phi_{zj}(x_2)\,, \quad f_{XZW}(x,z,w) = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} d_{zjk}\, \phi_{zj}(x)\, \phi_{zk}(z)\,,$$
$$g(x,z) = \sum_{j=1}^{\infty} b_{zj}\, \phi_{zj}(x)\,,$$

where the d_{zjk}'s and b_{zj}'s are generalised Fourier coefficients.

Let $r \geq 2p$ be an integer and put $\tau = 2r/(2r+q)$. We make the following assumptions, of which the first five are respectively analogous to A.1–A.5 in section 4.1. Let $C, \alpha, \beta > 0$.

MV.1. The data (X_i, W_i, Z_i, Y_i) are independent and identically distributed as (X, W, Z, Y), where X, W and Z are supported on $[0,1]^p$, $[0,1]^p$ and $[0,1]^q$, respectively, and $E\{Y - g(X,Z) \mid Z = z, W = w\} \equiv 0$.

MV.2. The distribution of (X, Z, W) has a density, f_{XZW}, with r derivatives of all types (when viewed as a function restricted to $[0,1]^{2p+q}$), each derivative bounded in absolute value by C; $g(x,z)$ and b_{zj} have r partial derivatives with respect to z, bounded in absolute value by C, uniformly in x and z; and the functions $E(Y^2 \mid Z = z, W = w)$ and $E(Y^2 \mid X = x, Z = z, W = w)$ are bounded uniformly by C.

MV.3. The constants α, β satisfy $\alpha > 1$ and $\alpha \le \beta\{(r-p)/(r+p)\} + \frac{1}{2}$, and moreover, $|b_{zj}| \le C\, j^{-\beta}$, $j^{-\alpha} \le C\, \lambda_{zj}$ and $\sum_{k \ge 1} |d_{zjk}| \le C\, j^{-\alpha/2}$, uniformly in $z \in [0,1]^q$, for all $j \ge 1$.

MV.4. The parameters a_n, h_x and h_z satisfy

$$a_n \asymp n^{-\alpha\tau/(2\beta+\alpha)}\,, \quad h \asymp n^{-2\beta\tau/\{(r+p)(2\beta+\alpha)\}}\,, \quad h_z \asymp n^{-1/(2r+q)}$$

as $n \to \infty$.

MV.5. The function $K_h(\,\cdot\,,\cdot\,)$ satisfies A.6 with, in place of (2.4),

$$\text{for all } t \in [0,1]\,, \quad h^{-(j+1)} \int_{t-1}^{t} u^j\, K_h(u,t)\, du = \begin{cases} 1 & \text{if } j = 0 \\ 0 & \text{if } 1 \le j \le r-1\,. \end{cases}$$

MV.6. For each $z \in [0,1]^q$ the functions ϕ_{zj} form an orthonormal basis for $L_2[0,1]^p$, and $\sup_x \sup_z \max_j |\phi_{zj}(x)| < \infty$.

Let $\mathcal{M} = \mathcal{M}(C, \alpha, \beta)$ denote the class of distributions of (X, W, Z, Y) that satisfy MV.1–MV.6.

Theorem 4.3. *As $n \to \infty$,*

$$\sup_{G \in \mathcal{M}} \sup_{z \in [0,1]^q} \int_{[0,1]^p} E_G\{\hat{g}(x,z) - g(x,z)\}^2\, dx = O\big(n^{-\tau(2\beta-1)/(2\beta+\alpha)}\big)\,.$$

The convergence rates here can be compared with those obtained by Darolles, Florens and Renault (2002), who treated the case $q = 0$. In their Theorem 4.2 they obtained:

$$\int_{[0,1]^p} E(\hat{g} - g)^2 = O\big(n^{-u}\big)\,,$$

where $u = \frac{1}{3}$ or $\frac{1}{2}$ under the respective pairs of conditions $\sum_j (b_j/\lambda_j)^2 < \infty$ & $p < 2r/3$, and $\sum_j (b_j^2/\lambda_j^4) < \infty$ & $p < r/2$. In terms of our assumptions, these

two pairs correspond to $\alpha < \beta - \frac{1}{2}$ and $\alpha < \frac{1}{2}\beta - \frac{1}{4}$, respectively, which imply that $\rho \equiv (2\beta-1)/(2\beta+\alpha) > \frac{1}{3}$ and $\rho > \frac{1}{2}$, respectively. Therefore, under MV1–MV6 the rate in given in our Theorem 4.3 is faster than that obtained by Darolles, Florens and Renault (2002).

4.4. Optimality. The convergence rates expressed by Theorems 4.1–4.3 are optimal in those contexts, in a minimax sense. Indeed, let $\tilde{g}$ denote any measurable functional of that data which is itself a measurable function on $[0,1]$ (in the cases of Theorems 4.1 and 4.2) or on $[0,1]^p$ (in the setting of Theorem 4.3); let $\mathcal{C}$ denote $\mathcal{G}$, $\mathcal{H}$ or $\mathcal{M}$ in the cases of Theorems 4.1–4.3, respectively; and put $\tau = 1$ in the contexts of Theorems 4.1 and 4.2, and $\tau = 2r/(2r+q)$ for Theorem 4.3.

Theorem 4.4.

$$\liminf_{n\to\infty} n^{\tau(2\beta-1)/(2\beta+\alpha)} \inf_{\tilde{g}} \sup_{G\in\mathcal{C}} \int E_G(\tilde{g}-g)^2 > 0\,. \tag{4.2}$$

In the multivariate setting of section 4.3 we interpret the integral at (4.2) as

$$\int_{[0,1]^p} E_G\{\tilde{g}(x,z) - g(x,z)\}^2\,dx\,,$$

and interpret Theorem 4.4 as stating that, for this representation, (4.2) holds for each $z \in [0,1]^q$.

5. Monte Carlo Experiments. This section reports the results of a Monte Carlo investigation of the finite-sample performance of the kernel estimator for the bivariate model. The estimator is described in section 2.3. Samples of size $n = 200$ were generated from the model determined by:

$$f_{XW}(x,w) = 2\,C_f \sum_{j=1}^{\infty} (-1)^{j+1}\, j^{-1}\, \sin(j\pi x)\, \sin(j\pi w)\,, \quad 0 \le x, w \le 1\,;$$

$$g(x) = 2^{1/2} \sum_{j=1}^{\infty} (-1)^{j+1}\, j^{-2}\, \sin(j\pi x)\,, \quad Y = E\{g(X) \mid W = w\} + V\,,$$

where C_f is a normalisation constant and V is distributed as Normal $\mathrm{N}(0, 0.01)$. For computational purposes the infinite series were truncated at $j = 100$. Figure 1 shows a graph of the marginal distributions of X and Z, which are identical. The solid line in Figure 2 depicts $g(x)$. The kernel function is the Epanechnikov kernel, $K(x) = 0.75\,(1-x^2)$ for $|x| \le 1$.

Each experiment consisted of estimating g at the 19 points, $x = 0.05, 0.10, \ldots,$ 0.95. The experiments were carried out in GAUSS using GAUSS pseudo-random number generators. There were 1000 Monte Carlo replications in each experiment.

Table 1 shows the performance of the estimator, $\hat{g}$, as a function of the bandwidth, h, and the ridge parameter, a_n. The quantities Bias^2, Var and MSE in the table were calculated as the averages, over the 19 values of x, of Monte Carlo approximations to pointwise squared bias, variance and mean squared error, respectively, at those points; the pointwise values were computed by averaging over the 1000 Monte Carlo simulations.

Results are illustrated graphically in Figure 2, for the case $h = 0.2$ and $a_n = 0.1$. The figure shows $g(x)$ (solid line), the Monte Carlo approximation to $E\{\hat{g}(x)\}$ (dashed line), and a 95% pointwise "estimation band." The band connects the points $g(x_j) \pm \delta_j$, for $j = 1, \ldots, 19$, where each δ_j is chosen so that the interval $[g(x_j) - \delta_j, g(x_j) + \delta_j]$ contains 95% of the 1000 simulated values of $\hat{g}(x_j)$. The figure shows, not surprisingly, that $\hat{g}$ is somewhat biased, but that the shape of $E\hat{g}$ is similar to that of g.

6. Technical arguments

6.1. Proof of Theorem 4.1. (The "big oh" bounds that we shall derive below apply uniformly in $G \in \mathcal{G}$, although for the sake of simplicity we shall not make this qualification.) Put $T^+ = (T + a_n)^{-1}$, let $\|\cdot\|$ denote the usual L_2 norm for functions from the interval $[0, 1]$ to the real line, and, given a functional χ from $L_2[0, 1]$ to itself, set

$$\|\chi\| = \sup_{\psi \in L_2[0,1] : \|\psi\| = 1} \|\chi(\psi)\| \,.$$

For future reference we note that A.3 and A.4 imply that

$$n^{\{1/(2\beta+\alpha)\}-1} a_n^{-1} + \left\{h^4 + h^{1/2} n^{-1} + (nh)^{-8/5}\right\} a_n^{-2} = O\left(n^{-(2\beta-1)/(2\beta+\alpha)}\right) . \quad (6.1)$$

Define

$$D_n(z) = \int g(x)\, f_{XW}(x, w)\, T^+(\hat{f}_{XW} - f_{XW})(z, w)\, dx\, dw \,,$$

$$A_{n1}(z) = \frac{1}{n} \sum_{i=1}^{n} \left(T^+ f_{XW}\right)(z, W_i)\, Y_i \,,$$

$$A_{n2}(z) = \frac{1}{n} \sum_{i=1}^{n} \{T^+(\hat{f}_{XW}^{(-i)} - f_{XW})\}(z, W_i)\, Y_i - D_n(z)\,,$$

$$A_{n3}(z) = \frac{1}{n} \sum_{i=1}^{n} \{(\widehat{T}^+ - T^+) f_{XW}\}(z, W_i)\, Y_i + D_n(z)\,,$$

$$A_{n4}(z) = \frac{1}{n} \sum_{i=1}^{n} \{(\widehat{T}^+ - T^+)\,(\hat{f}_{XW}^{(-i)} - f_{XW})\}(z, W_i)\, Y_i\,.$$

Then, $\hat{g} = A_{n1} + \ldots + A_{n4}$, and so the theorem will follow if we prove that

$$E\|A_{n1} - g\|^2 = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big)\,, \tag{6.2}$$

$$E\|A_{nj}\|^2 = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big) \quad \text{for } j = 2, 3, 4\,. \tag{6.3}$$

To derive (6.2), note that $EA_{n1} - g = -a_n \sum_{j\ge1} b_j\,(\lambda_j + a_n)^{-1}\,\phi_j$. Therefore,

$$\|EA_{n1} - g\|^2 = a_n^2 \sum_{j=1}^{\infty} \frac{b_j^2}{(\lambda_j + a_n)^2}\,.$$

Divide the last-written series up into the sum over $j \le J \equiv a_n^{-1/\alpha}$, and the complementary part, thereby bounding the right-hand side by $a_n^2 \sum_{j\le J} (b_j/\lambda_j)^2 + \sum_{j>J} b_j^2$; and use A.3 and A.4 to bound each of these terms; hence proving that

$$\|EA_{n1} - g\|^2 = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big)\,. \tag{6.4}$$

Using A.2 we deduce that

$$\begin{aligned} n\,\mathrm{var}\{A_{n1}(z)\} &\le E\big[\{(T^+ f_{XW})(z, W)\, Y\}^2\big] = E\big[\{(T^+ f_{XW})(z, W)\}^2\, E\big(Y^2 \,\big|\, W\big)\big] \\ &\le \text{const.}\, E\big[\{(T^+ f_{XW})(z, W)\}^2\big]\,, \end{aligned}$$

where, here and below, "const." will denote a positive constant, different at different appearances. Let t^+ denote the kernel of the transformation T^+, and write ψ_z for the function $\psi_z(x_1) = t^+(x_1, z)$. Then, $n \int \mathrm{var}\{A_{n1}(z)\}\,dz$ is dominated by a constant multiple of

$$\begin{aligned} &\int E\big[\{(T^+ f_{XW})(z, W)\}^2\big]\,dz \\ &\qquad = \iiint t^+(x_1, z)\, t^+(x_2, z)\, f_{XW}(x_1, w)\, f_{XW}(x_2, w)\, f_W(w)\,dw \\ &\qquad \le \text{const.} \iiint t^+(x_1, z)\, t(x_1, x_2)\, t^+(x_2, z)\,dx_1\,x_2\,dz \\ &\qquad = \text{const.} \int (T^+ T \psi_z)(z)\,dz\,, \end{aligned}$$

where all integrals are over $[0,1]$. Since T^+T is a bounded, positive-definite operator, and since $0 \le \psi_z(x) \le \text{const.}\, a_n^{-1}$ for all x and z, then $\int (T^+T\psi_z)(z)\,dz = O(a_n^{-1})$ as $n \to \infty$. Therefore, $n \int \text{var}\{A_{n1}(z)\}\,dz = O(a_n^{-1})$, and so,

$$E\|A_{n1} - EA_{n1}\|^2 = O\big(n^{-1}a_n^{-1}\big) = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big)\,.$$

Result (6.2) is implied by this bound and (6.4).

Next we derive (6.3) in the case $j = 2$. Here and below, given a bivariate function $\phi(z,w)$, put $\phi_w(z) = \phi(z,w)$ and define $T^+\phi(z,w) = (T^+\phi_w)(z)$. Let

$$\begin{aligned}
D_{ni}(z) &= \int g(x)\, f_{XW}(x,w)\, T^+\big(\hat f_{XW}^{(-i)} - f_{XW}\big)(z,w)\,dx\,dw\,,\\
A_{n21}(z) &= \frac{1}{n}\sum_{i=1}^n \Big\{T^+\big(\hat f_{XW}^{(-i)} - f_{XW}\big)(z,W_i)\,Y_i - D_{ni}(z)\Big\}\,,\\
A_{n22}(z) &= \frac{1}{n}\sum_{i=1}^n \{D_{ni}(z) - D_n(z)\}\,,
\end{aligned}$$

in which notation $A_{n2} = A_{n21} + A_{n22}$. Write $\int A_{n21}(z)^2\,dz$ as a double series, and take the expected values of the terms one by one. It may be shown by tedious calculation that the total contribution of the terms equals $O\{h^4\,(na_n^2)^{-1} + (nha_n)^{-2}\}$. Therefore,

$$E\|A_{n21}\|^2 = O\big\{h^4\,(na_n^2)^{-1} + (nha_n)^{-2}\big\} = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big)\,, \tag{6.5}$$

where we used (6.1) to obtain the second identity. Furthermore,

$$A_{n22}(z) = -n^{-1}\int g(x)\, f_{XW}(x,w)\, T^+\hat f_{XW}(z,w)\,dx\,dw\,,$$

from which, noting (6.1), it may be deduced that

$$\begin{aligned}
E\|A_{n22}\|^2 &\le \text{const.}\,(na_n)^{-2}\,E\bigg(\int \big|g\, f_{XW}\,\hat f\big|\bigg)^2\\
&= O\big\{(na_n)^{-2}\big\} = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big)\,.
\end{aligned}$$

Property (6.3), in the case $j = 2$, follows from this result and (6.5).

Next we derive (6.3) for $j = 3$. Define

$$A_{n31} = -(I + T^+\Delta)^{-1}T^+\Delta g + D_n\,, \quad A_{n32} = -(I + T^+\Delta)^{-1}T^+\Delta(A_{n1} - g)\,.$$

Noting that $\widehat T^+ - T^+ = -(I + T^+\Delta)^{-1}T^+\Delta T^+$, it can be seen that $A_{n3} = A_{n31} + A_{n32}$.

Let $\delta = h^4 + (nh)^{-1}$ and $\Delta = \widehat{T} - T$, the latter an operator. Using standard, but tedious, moment calculations it may be proved that $E(\hat{t}-t)^{2k} = O(\delta^k)$ for each integer $k \geq 1$, uniformly in the argument of $\hat{t} - t$. (The quantity δ involves $(nh)^{-1}$, rather than $(nh^2)^{-1}$, since the integral in the definition of $\hat{t}$ effectively removes one of the factors h^{-1}.) Therefore, since $\|\Delta\|^2 = \int(\hat{t}-t)^2$, then for each integer $k \geq 1$,

$$E\|\Delta\|^{2k} = O\big(\delta^k\big)\,. \tag{6.6}$$

At the end of this proof we shall show that for each $k \geq 1$,

$$E\big\{\|(I + T^+\Delta)^{-1}\|^k\big\} = O(1) \tag{6.7}$$

as $n \to \infty$. Hence, using the Cauchy-Schwarz inequality,

$$\begin{aligned}\big\{E\|(I + T^+\Delta)^{-1}T^+\Delta\|^4\big\}^2 &\leq E\|(I + T^+\Delta)^{-1}\|^8\,\|T^+\|^8\,E\|\Delta\|^8 \\ &= O\big(\delta^4/a_n^8\big)\,.\end{aligned} \tag{6.8}$$

From this result, and the Cauchy-Schwarz inequality again, we obtain:

$$\begin{aligned}E\|A_{n32}\|^2 &\leq \big\{E\|(I + T^+\Delta)^{-1}T^+\Delta\|^4\,E\|A_{n1} - g\|^4\big\}^{1/2} \\ &= O\big\{(\delta/a_n^2)^2\,\big(E\|A_{n1} - g\|^4\big)^{1/2}\big\} = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big)\,,\end{aligned} \tag{6.9}$$

the final identity following using an argument similar to that leading to (6.2).

Put

$$\begin{aligned}B_{n1}(z) &= \int\{\hat{f}_{XW}(x,w) - f_{XW}(x,w)\}\,f_{XW}(z,w)\,g(x)\,dx\,dw\,, \\ B_{n2}(z) &= \int\{\hat{f}_{XW}(z,w) - f_{XW}(z,w)\}\,f_{XW}(x,w)\,g(x)\,dx\,dw\,, \\ B_{n3}(z) &= \int\{\hat{f}_{XW}(x,w) - f_{XW}(x,w)\}\,\{\hat{f}_{XW}(z,w) - f_{XW}(x,w)\}\,g(x)\,dx\,dw\,, \\ B_{n11}(z) &= \int\{E\hat{f}_{XW}(x,w) - f_{XW}(x,w)\}\,f_{XW}(z,w)\,g(x)\,dx\,dw\,, \\ B_{n12}(z) &= \int\{\hat{f}_{XW}(x,w) - E\hat{f}_{XW}(x,w)\}\,f_{XW}(z,w)\,g(x)\,dx\,dw\,, \\ B_{n21}(z) &= \int\{E\hat{f}_{XW}(z,w) - f_{XW}(z,w)\}\,f_{XW}(x,w)\,g(x)\,dx\,dw\,, \\ B_{n22}(z) &= \int\{\hat{f}_{XW}(z,w) - E\hat{f}_{XW}(z,w)\}\,f_{XW}(x,w)\,g(x)\,dx\,dw\,.\end{aligned}$$

In this notation, $\Delta g = B_{n1} + B_{n2} + B_{n3}$, $B_{n1} = B_{n11} + B_{n12}$, $B_{n2} = B_{n21} + B_{n22}$ and $T^+ B_{n2} = D_n$, whence

$$A_{n31} = -(I + T^+\Delta)^{-1} T^+ (B_{n11} + B_{n12} + B_{n3}) \\ + (I + T^+\Delta)^{-1} T^+ \Delta T^+ (B_{n21} + B_{n22}) \,.$$

From this property, (6.7) and the Cauchy-Schwarz inequality we deduce that

$$E\|A_{n31}\|^2 \le \text{const.} \Big(\|T^+ B_{n11}\|^4 + E\|T^+ B_{n12}\|^4 + E\|T^+ \Delta T^+ B_{n21}\|^4 \\ + E\|T^+ \Delta T^+ B_{n22}\|^4 + E\|T^+ B_{n3}\|^4 \Big)^{1/2} . \quad (6.10)$$

Since $\|B_{n11}\| + \|B_{n21}\| = O(h^2)$ and $\|T^+\| = O(a_n^{-1})$ then, by (6.1),

$$\|T^+ B_{n11}\| + \|T^+ B_{n21}\| \le \|T^+\| \, (\|B_{n11}\| + \|B_{n21}\|) \\ = O\big(h^2 a_n^{-1}\big) = O\big(n^{-(2\beta-1)/\{2(2\beta+\alpha)\}}\big) \,. \quad (6.11)$$

Furthermore, with

$$\Delta_{jk} = \int \{\hat f_{XW}(x,w) - E\hat f_{XW}(x,w)\} \, \phi_j(x) \, \phi_k(x) \, dx \, dw \,,$$

we have

$$T^+ B_{n12}(z) = \sum_{j=1}^\infty \sum_{k=1}^\infty \sum_{\ell=1}^\infty \frac{d_{jk} \, b_\ell \, \Delta_{\ell k}}{\lambda_j + a_n} \, \phi_j(z) \,.$$

Now, $E(\Delta_{j_1 k_1} \, \Delta_{\ell_1 m_1} \, \Delta_{j_2 k_2} \, \Delta_{\ell_2 m_2}) = O(n^{-2})$, uniformly in the indicated indices; $\sum_\ell |b_\ell| < \infty$, since A.3 implies that $\beta > 1$; and $\sum_{k \ge 1} |d_{jk}| = O(j^{-\alpha/2})$, again by A.3. Therefore,

$$\big(E\|T^+ B_{n12}\|^4\big)^{1/2} = \left[E\Big\{ \sum_{j=1}^\infty \frac{1}{(\lambda_j + a_n)^2} \Big(\sum_{k=1}^\infty \sum_{\ell=1}^\infty d_{jk} \, b_\ell \, \Delta_{\ell k} \Big)^2 \Big\}^2 \right]^{1/2} \\ = O\left\{ \frac{1}{n} \sum_{j=1}^\infty \frac{1}{(\lambda_j + a_n)^2} \Big(\sum_{k=1}^\infty \sum_{\ell=1}^\infty |d_{jk}| \, |b_\ell| \Big)^2 \right\} \\ = O\left\{ \frac{1}{n} \sum_{j=1}^\infty \frac{j^{-\alpha}}{(\lambda_j + a_n)^2} \right\} = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big) \,. \quad (6.12)$$

Note too that

$$E\|T^+ B_{n22}\|^8 \le \|T^+\|^8 \, E\|B_{n22}\|^8 = O\big(a_n^{-4} \, E\|B_{n22}\|^8\big) \\ = O\left\{ a_n^{-4} \, E\Big(\int B_{n22}^2 \Big)^4 \right\} = O\big\{(nha_n^2)^{-4}\big\} \,, \quad (6.13)$$

the last inequality following by moment calculations similar to those leading to (6.6). In view of (6.1) and (6.6),

$$E\|T^+\Delta\|^8 \le \|T^+\|^8\, E\|\Delta\|^8 = O\big(a_n^{-8}\, E\|\Delta\|^8\big) = O\big(\delta^4/a_n^8\big) = O(1)\,. \tag{6.14}$$

By (6.11), (6.13), (6.14) and the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\big(E\|T^+\Delta T^+ B_{n21}\|^4\big)^{1/2} + \big(E\|T^+\Delta T^+ B_{n22}\|^4\big)^{1/2} \\
&\qquad \le \big(E\|T^+\Delta\|^4\, \|T^+ B_{n21}\|^4\big)^{1/2} + \big(E\|T^+\Delta\|^8\, E\|T^+ B_{n22}\|^8\big)^{1/4} \\
&\qquad = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big) + O\Big\{\big(\delta/a_n^2\big)\,\big(nha_n^2\big)^{-1}\Big\} \\
&\qquad = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big)\,;
\end{aligned} \tag{6.15}$$

we used (6.1) to obtain the last identity.

Define

$$\begin{aligned}
I_n(w) &= \int \{\hat f_{XW}(x,w) - f_{XW}(x,w)\}\, g(x)\, dx\,, \\
J_n &= \iint \big\{T^+(\hat f_{XW} - f_{XW})(z,w)\big\}^2\, dw\, dz\,.
\end{aligned}$$

Moment calculations show that $E\|I_n\|^8 = O(\delta^4)$ and $E(J_n^4) = O(\delta^4/a_n^8)$, and so by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\big(E\|T^+ B_{n3}\|^4\big)^{1/2} &\le \big\{E\big(\|I_n\|^4\, J_n^2\big)\big\}^{1/2} \le \big(E\|I_n\|^8\, EJ_n^4\big)^{1/4} \\
&= O(\delta^2/a_n^2) = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big)\,.
\end{aligned} \tag{6.16}$$

Result (6.3), for $j = 3$, follows from (6.9)–(6.12), (6.15) and (6.16).

Next we derive (6.3) for $j = 4$. Since $\widehat T^+ - T^+ = -(I + T^+\Delta)^{-1}T^+\Delta T^+$ and $I - \widehat T^+ T = -(I + T^+\Delta)^{-1}T^+\Delta$ then

$$A_{n4} = -(I + T^+\Delta)^{-1}T^+\Delta(A_{n2} - T^+ B_{n2})\,.$$

The arguments leading to (6.3) with $j = 2$, and (6.15), may be used to prove that

$$\eta \equiv \Big\{(\delta^2/a_n)^4\, E\|A_{n2}\|^4 + E\|T^+\Delta T^+ B_{n2}\|^4\Big\}^{1/2} = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big)\,.$$

Therefore, by (6.7), (6.8) and the Cauchy-Schwarz inequality,

$$\begin{aligned}
E\|A_{n4}\|^2 &\le 2\Big\{E\|(I + T^+\Delta)^{-1}T^+\Delta\|^4\, E\|A_{n2}\|^4\Big\}^{1/2} \\
&\qquad + 2\Big\{E\|(I + T^+\Delta)^{-1}\|^4\, E\|T^+\Delta T^+ B_{n2}\|^4\Big\}^{1/2} \\
&= O(\eta) = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big)\,.
\end{aligned}$$

This proves (6.3) for $j = 4$.

It remains to derive (6.7). Let $\psi \in L_2[0,1]$. Then, for constants not depending on ψ, if $\|T^+\Delta\| \le \frac{1}{2}$,

$$\|(I + T^+\Delta)^{-1}\psi\| \le \text{const.}\, \|\psi\|\,,$$

and, without any constraint on $\|T^+\Delta\|$,

$$\|(I + T^+\Delta)^{-1}\psi\| = \|\widehat{T}^+(T + a_n)\psi\| \le \|\widehat{T}^+\|\,\|T + a_n\|\,\|\psi\| \le \text{const.}\, a_n^{-1}\,\|\psi\|\,.$$

Therefore,

$$\|(I + T^+\Delta)^{-1}\| \le \text{const.}\, \{1 + a_n^{-1}\, I(\|T^+\Delta\| > \tfrac{1}{2})\}\,.$$

Hence, noting (6.6), and employing Markov's inequality to bound $P(\|T^+\Delta\| > \frac{1}{2})$, we deduce that for each fixed $k, \ell > 0$,

$$\begin{aligned} E\{\|(I + T^+\Delta)^{-1}\|^k\} &\le \text{const.}\, \{1 + a_n^{-k}\, P(\|T^+\Delta\| > \tfrac{1}{2})\} \\ &\le \text{const.}\, \{1 + a_n^{-k}\, E(\|T^+\Delta\|^{2\ell})\} \\ &\le \text{const.}\, \{1 + a_n^{-k-2\ell}\, E(\|\Delta\|^{2\ell})\} \\ &\le \text{const.}\, (1 + a_n^{-k-2\ell}\, \delta^\ell) = \text{const.}\, \{1 + a_n^{-k}\, (\delta/a_n^2)^\ell\}\,, \end{aligned} \tag{6.17}$$

where the constants depend on k and ℓ but not on n. If k is given then we may choose $\ell = \ell(k)$ so large that $a_n^{-k}\,(\delta/a_n^2)^\ell \to 0$ as $n \to \infty$, and so (6.7) follows from (6.17).

6.2. Proof of Theorem 4.2. Put $\bar{p} = (p_1, \ldots, p_m)^{\mathrm{T}}$, where $p_j = E_G\{g(\widetilde{X})\,\chi_j(\widetilde{W})\} = E_G\{Y\chi_j(\widetilde{W})\}$. Let $\gamma = (\gamma_j)$ and $p = (p_j)$ denote infinite column vectors, and let $\bar{Q}$ be the $m \times m$ upper left-hand sub-matrix of Q. Since $p = Q\gamma$ then $p_j = p_j(G) = O(j^{-(2\beta+\alpha)/2})$, uniformly in $G \in \mathcal{H}$, as $j \to \infty$. Therefore, $(\bar{Q}^{\mathrm{T}}p)_i = O(i^{-(\alpha+\beta)})$, uniformly in $1 \le i \le m$, $n \ge 1$ and $G \in \mathcal{H}$. This result will be used below without further reference.

Put $\bar{M} = \bar{Q}\,\bar{Q}^{\mathrm{T}} + a_n I_m$ and $\widehat{M} = \widehat{Q}\widehat{Q}^{\mathrm{T}} + a_n I_m$. It may be deduced from the definition of $\mathcal{H}$ that the bounds on $|q_{jk}|$ and $|q_{jk}^{(-1)}|$ in that definition apply too to the (j,k)th elements of $\bar{M}$ and $\bar{M}^{-1}$, respectively, provided we replace α by 2α and alter the constants C_1 and C_2 (retaining their positivity, of course). The bounds are valid uniformly in $1 \le j, k \le m$ and $n \ge 1$, and permit it to be proved that

$$\bigl(\bar{M}^{-1}\,\bar{Q}^{\mathrm{T}}\bar{p}\bigr)_j = \bigl\{\bigl(Q^{\mathrm{T}}Q\bigr)^{-1}Q^{\mathrm{T}}p\bigr\}_j + O\bigl(m^{-\beta}\bigr) = \gamma_j + O\bigl(m^{-\beta}\bigr)\,,$$

uniformly in $1 \le i \le m$, $n \ge 1$ and distributions of $G \in \mathcal{H}$. Note too that

$$\begin{aligned}\widehat{M}^{-1} &= \{I + \bar{M}^{-1}(\widehat{M} - \bar{M})\}^{-1}\bar{M}^{-1}\,,\\ \widehat{M}^{-1}\widehat{Q}^{\mathrm{T}}\hat{p} - \bar{M}^{-1}\bar{Q}^{\mathrm{T}}\bar{p} &= \{\bar{M}^{-1} + (\widehat{M}^{-1} - \bar{M}^{-1})\}\,(\widehat{Q}^{\mathrm{T}}\hat{p} - \bar{Q}^{\mathrm{T}}\bar{p})\\ &\quad + (\widehat{M}^{-1} - \bar{M}^{-1})\,\bar{Q}^{\mathrm{T}}\bar{p}\,.\end{aligned}$$

From these properties it may be shown that

$$\begin{aligned}E_G\Big\{\sum_{j=1}^{m}(\tilde{\gamma}_j - \gamma_j)^2\Big\} = O\Big\{&E_G\Big(\sum_{i=1}^{m}\big[\{\bar{M}^{-1}\,(\widehat{Q}^{\mathrm{T}}\hat{p} - \bar{Q}^{\mathrm{T}}\bar{p})\}_j\big]^2\Big)\\ &+ E_G\Big(\sum_{i=1}^{m}\big[\{\bar{M}^{-1}(\widehat{M} - \bar{M})\,\bar{M}^{-1}(\widehat{Q}^{\mathrm{T}}\hat{p} - \bar{Q}^{\mathrm{T}}\bar{p})\}_j\big]^2\Big)\\ &+ E_G\Big(\sum_{i=1}^{m}\big[\{\bar{M}^{-1}(\widehat{M} - \bar{M})\,\bar{M}^{-1}\,\bar{Q}^{\mathrm{T}}\bar{p}\}_j\big]^2\Big) + m^{1-2\beta}\Big\}\,,\end{aligned} \tag{6.18}$$

uniformly in $G \in \mathcal{H}$.

It may be proved by Taylor expansion arguments, involving approximating $\widehat{W}_i = \widehat{F}_W(W_i)$ by $\widetilde{W}_i = F_W(W_i)$, and analogously for $\widehat{X}_i$ and $\widetilde{X}_i$, that for each $r, \epsilon > 0$,

$$\max_{1\le j,k\le n^{(1/2)-\epsilon}}\ \sup_{G\in\mathcal{H}}\ E_G|\hat{q}_{jk} - q_{jk}|^r = O\big(n^{-r/2}\big)\,. \tag{6.19}$$

$$\max_{1\le j\le n^{(1/2)-\epsilon}}\ \sup_{G\in\mathcal{H}}\ E_G\big(\hat{p}_j - p_j\big)^2 = O\big(n^{-1}\big)\,. \tag{6.20}$$

Rather standard, but tedious, moment calculations, using (6.19) and (6.20), may be employed to show that each of the expected values on the right-hand side of (6.18) equals $O(n^{-1}m^{\alpha+1})$, uniformly in $G \in \mathcal{H}$. Therefore,

$$\sup_{G\in\mathcal{H}}\sum_{j=1}^{m}E_G\big\{(\tilde{\gamma}_j - \gamma_j)^2\big\} = O\big(n^{-1}m^{\alpha+1} + m^{1-2\beta}\big) = O\big(n^{-(2\beta-1)/(2\beta+\alpha))}\big)\,. \tag{6.21}$$

It follows from the definition of $\mathcal{H}$ that $\sum_{j>m}\gamma_j^2 = O(m^{1-2\beta})$, uniformly in $G \in \mathcal{H}$. This result, and (6.21), imply that

$$\int E_G(\bar{g} - g)^2 = \sum_{j=1}^{m}E_G(\tilde{\gamma}_j - \gamma_j)^2 + \sum_{j=m+1}^{\infty}\gamma_j^2 = O\big(n^{-(2\beta-1)/(2\beta+\alpha)}\big)\,,$$

uniformly in $G \in \mathcal{H}$, completing the proof of the theorem.

6.3. Proof of Theorem 4.4. For simplicity we deal only with the kernel and orthogonal series settings, discussed in sections 4.1 and 4.2, respectively, where Z is not present. We may assume the following: $\phi_j \equiv \chi_j$, $\phi_1 \equiv 1$ and $\phi_{j+1}(x) = 2^{-1/2} \cos(j\pi x)$, for $j \geq 1$; the marginal distributions of X and W are uniform on the unit interval; and

$$f_{XW}(x,w) = \sum_{j=1}^{\infty} j^{-\alpha/2}\, \phi_j(x)\, \phi_j(w)\,, \quad Y = \sum_{j=m+1}^{2m} \theta_j\, j^{-(2\beta+\alpha)/2}\, \phi_j(W) + V\,, \tag{6.22}$$

where m equals the integer part of $n^{1/(2\beta+\alpha)}$, the θ_j's are all either 0 or 1, and V is Normal $\mathrm{N}(0,1)$, independent of (X,W).

The function g implied by (6.22) is $g(x) = \sum_{m+1\leq j\leq 2m} \theta_j\, j^{-\beta}\, \phi_j(x)$. Note too that if $\tilde{g}$ is an estimator of g then

$$\tilde{\theta}_j = j^{\beta} \int \tilde{g}\, \phi_j \tag{6.23}$$

may be viewed as an estimator of θ_j.

A standard argument based on the Neyman-Pearson lemma shows that

$$\liminf_{n\to\infty} \inf_{m+1\leq j\leq 2m} \inf_{\check{\theta}_j} \sup{}^{*}\ E(\check{\theta}_j - \theta_j)^2 > 0\,,$$

where $\sup^*$ denotes the supremum over all 2^m different distributions of (X,W,Y) obtained by taking different choices of $\theta_{m+1},\ldots,\theta_{2m}$ in (6.22), and $\inf_{\check{\theta}_j}$ represents the infimum over all measurable functions $\check{\theta}_j$ of the data. Therefore, if $\tilde{g}$ is given, and $\tilde{\theta}_{m+1},\ldots,\tilde{\theta}_{2m}$ are the estimators of $\theta_{m+1},\ldots,\theta_{2m}$, respectively, derived from $\tilde{g}$ as suggested at (6.23), then

$$\begin{aligned} \sup{}^{*} \int (\tilde{g}-g)^2 &= \sup{}^{*} \sum_{j=m+1}^{2m} E(\tilde{\theta}_j - \theta_j)^2\, j^{-2\beta} \\ &\geq \text{const.} \sum_{j=m+1}^{2m} j^{-2\beta} \geq \text{const.}\, j^{-(2\beta-1)/(2\beta+\alpha)}\,, \end{aligned}$$

where the constants do not depend on choice of $\tilde{g}$. This proves the theorem.

REFERENCES

BLUNDELL, R. AND POWELL, J.L. (2003). Endogeneity in nonparametric and semiparametric regression models. In: *Advances in Economics and Econometrics: Theory and Applications*, Dewatripont, M., Hansen, L.-P. and

Turnovsky, S.J., eds, vol. 2, pp. 312–357. Cambridge, UK: Cambridge University Press.

DAROLLES, S., FLORENS, J.-P. AND RENAULT, E. (2002). Nonparametric instrumental regression. Working paper, GREMAQ, University of Social Science, Toulouse.

FLORENS, J.P. (2003). Inverse problems and structural econometrics: the example of instrumental variables. In: *Advances in Economics and Econometrics: Theory and Applications*, Dewatripont, M., Hansen, L.-P. and Turnovsky, S.J., eds, vol. 2, pp. 284–311. Cambridge, UK: Cambridge University Press.

GROETSCH, C. (1984). *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. London: Pitman.

KRESS, R. (1999). *Linear Integral Equations*, 2nd Edn. New York: Springer.

NASHED, M.Z. AND WAHBA, G. (1984). Generalized inverse in reproducing kernel spaces: an approach to regularization of linear operator equations. *SIAM J. Math. Anal.* **5**, 974-987.

NEWEY, W.K. AND POWELL, J.L. (2002). Instrumental variable estimation of nonparametric models. *Econometrica*, to appear.

NEWEY, W.K., POWELL, J.L. AND VELLA, F. (1999). *Econometrica* **67**, 565–603.

O'SULLIVAN, F. (1986). A statistical perspective on ill-posed problems. *Statist. Sci.* **1**, 502–527.

TIKHONOV, A. AND ARSENIN, B. (1977). *Solutions of Ill-Posed Problems*. Washington, D.C.: Winston.

VAN ROOIJ, O. AND RUYMGAART, C.H. (1999). On inverse estimation. In: *Asymptotics, Nonparametrics, and Time Series*, S. Ghosh, ed., pp. 579-613. New York: Dekker.

WAHBA, G. (1973). Convergence rates of certain approximate solutions of Fredholm integral equations of the first kind, *J. Approximation Theory* **7**, 167-185.

Captions for table and figures.

Table 1. Results of Monte Carlo Experiments

Figure 1. Density of X and W used in Monte Carlo Experiments.

Figure 2. Graph of 95% Estimation Band. The solid, dashed and dotted lines show g, $E(\hat{g})$ and the 95% estimation band, respectively.

Table 1: Results of Monte Carlo Experiments

a_n	h	Bias^2	Var	MSE
0.05	0.10	0.0039	0.0321	0.0361
	0.20	0.0065	0.0162	0.0227
	0.30	0.0262	0.0119	0.0381
	0.40	0.0525	0.0087	0.0612
0.10	0.10	0.0118	0.0221	0.0339
	0.20	0.0105	0.0115	0.0215
	0.30	0.0141	0.0078	0.0219
	0.40	0.0263	0.0062	0.0325
0.15	0.10	0.0224	0.0190	0.0414
	0.20	0.0165	0.0098	0.0263
	0.30	0.0149	0.0063	0.0212
	0.40	0.0220	0.0049	0.0269
0.20	0.10	0.0335	0.0174	0.0508
	0.20	0.0268	0.0081	0.0349
	0.30	0.0214	0.0058	0.0272
	0.40	0.0252	0.0044	0.0295

Figure 1: Density of X and W Used in Monte Carlo Experiments

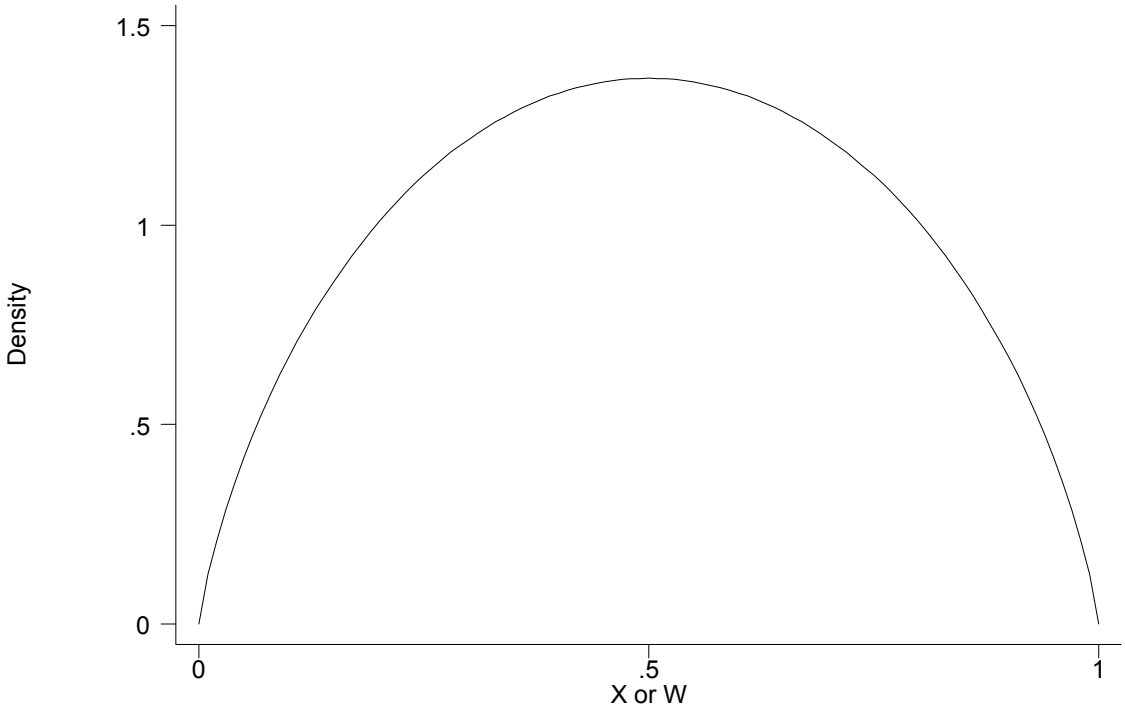


Figure 2: Graph of 95% Estimation Band. The solid, dashed, and dotted lines show g , $E\hat{g}$, and the 95% estimation band, respectively