Econometric Theory, **27**, 2011, 609–638.
doi:10.1017/S0266466610000423

DEMAND ANALYSIS AS AN ILL-POSED INVERSE PROBLEM WITH SEMIPARAMETRIC SPECIFICATION

STEFAN HODERLEIN
Brown University

HAJO HOLZMANN
Marburg University

In this paper we are concerned with analyzing the behavior of a semiparametric estimator that corrects for endogeneity in a nonparametric regression by assuming mean independence of residuals from instruments only. Because it is common in many applications, we focus on the case where endogenous regressors and additional instruments are jointly normal, conditional on exogenous regressors. This leads to a severely ill-posed inverse problem. In this setup, we show first how to test for conditional normality. More importantly, we then establish how to exploit this knowledge when constructing an estimator, and we derive the large sample behavior of such an estimator. In addition, in a Monte Carlo experiment we analyze its finite sample behavior. Our application comes from consumer demand. We obtain new and interesting findings that highlight both the advantages and the difficulties of an approach that leads to ill-posed inverse problems. Finally, we discuss the somewhat problematic relationship between endogenous nonparametric regression models and the recently emphasized issue of unobserved heterogeneity in structural models.

1. INTRODUCTION

1.1. General Motivation

Regression models with endogenous regressors are commonplace throughout much of applied economics, and are perhaps the rule rather than the exception. To deal with this problem, most of the time applied researchers employ instrumental variables (in the following denoted W). While it is well established how they are to be used if the model is known to be linear, it is much less clear in more general settings. Indeed, in this paper we will be concerned with the most general, the nonparametric setting, where it is only known that the structural relationship between a variable Y and an endogenous regressor X is a smooth but otherwise

The authors would like to thank Jean-Pierre Florens, Peter C.B. Phillips, and two anonymous referees for their helpful comments. Financial support of the Landesstiftung Baden-Württemberg "Eliteförderungsprogramm," the DFG, grant HO 3260/3-1, the Landesstiftung Baden-Württemberg, Juniorprofessorenprogramm, and the Claussen-Simon-Stiftung is gratefully acknowledged. Address correspondence to Stefan Hoderlein, Department of Economics, Box B, Providence, RI 02912, USA; e-mail: stefan_hoderlein@yahoo.com.

© Cambridge University Press 2010

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

unrestricted function m. In this setting we focus specifically on the case when the endogenous regressor X and the instrument W are jointly normally distributed.

This choice of focus is intentional: Indeed, the aim of the paper is to show the scope of the nonparametric intrumental variable (IV) approach in a typical economic application — and (approximate) normality is arguably the leading distribution of continuous variables we find in applications.[1] Consider for instance Figure 1, which shows a kernel density estimate of the marginal distribution of log total expenditure against a maximum likelihood estimate of a normal distribution using the same data. Judging from the marginals, normality of the joint distribution of log income and total expenditure seems to provide an acceptable description of the data.

In consumer demand, a well-developed area of application, log total expenditure is suggested by economic theory as one of the regressor. However, it is believed to be endogenous, and log disposable income is usually taken as instrument. As we will demonstrate below, both variables are well characterized by joint normality, conditional on household covariables. While normality is commonly associated with benign behavior, this is not at all the case in a nonparametric regression with endogenous regressors. Indeed, joint normality can lead to very slow rates of convergence, as demonstrated by various authors (Blundell, Chen, and Kristensen, 2007 (BCK); Darolles, Florens, and Renault, 2004; and Severini and Tripathi, 2006).

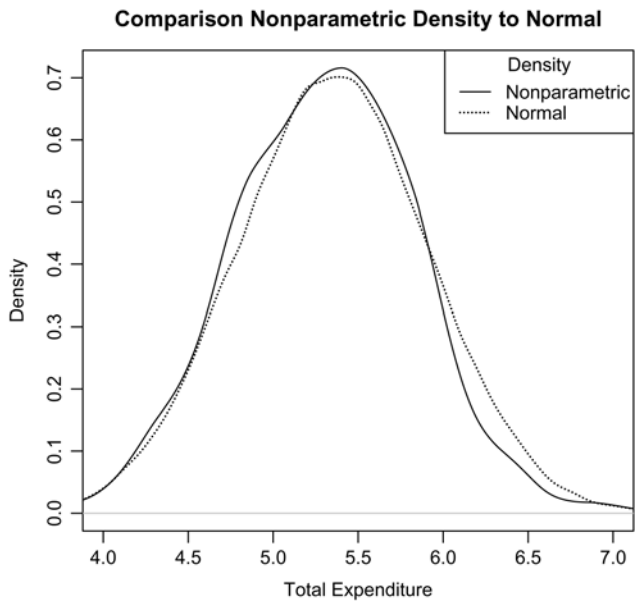


FIGURE 1. Marginal density comparison.

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

In this paper we examine what can be learned about the relationship of interest m in such a scenario, and by which means. Specifically, we consider the population model

$$Y = m(X, Z) + U, \qquad \mathbb{E}(U|W, Z) = 0, \tag{1.1}$$

where m is an unknown function; Y, X, and W are as defined above; Z denotes additional exogenous regressors; and U is an (unobservable) random error. This model will be called a nonparametric IV regression model.

We propose a series estimator in the Hermite polynomials where the coefficients are obtained by additional nonparametric regression of the relevant quantities on Z. In comparison with other estimators in related settings, our estimator has the main advantages its construction is intuitive, it is simple and transparent, it is fairly easy to compute, and it is designed for an economically particularly relevant situation (model (1.1) with continuously distributed Z and conditional normality of (X, W) given Z). Given the complexity of the other models in this area, for these reasons it should be rather attractive for applied researchers.

1.2. Relationship to the Literature

The approach put forward in this paper is nonparametric, but under a semiparametric specification. Moreover, it is motivated by an application and also provides, to the best of our knowledge, the first application of a kernel-based nonparametric model to an economic application in an ill-posed scenario. Therefore, it is most closely related to the model of BCK, who successfully implement a generalized partially linear model using the method of sieves. We compare our approach to theirs at various places throughout the paper. Specifically, we discuss specification issues arising from heterogeneity of a population of rational individuals for both approaches in Section 5. Here we just mention that BCK assume discrete exogenous covariates Z and model their influence semiparametrically. For the nonparametric part, BCK use a sieve estimator. As candidate sieve spaces, they consider the Hermite functions (not the polynomials), but in addition other bases like spline bases. Thus, their estimation strategy allows for quite some flexibility by allowing them to choose the appropriate sieves. They introduce a sieve measure of ill-posedness, and obtain rates of convergence both for the mildly ill-posed (polynomial decay) as well as severely ill-posed (exponential decay) case.

Other than through the (testable) semiparametric element of joint normality, our paper is entirely nonparametric. Fully nonparametric estimation of m in model (1.1) has been studied quite intensively in recent years. Here we put our assumptions and results into perspective with the existing literature, specifically with Carrasco, Florens, and Renault (2005), Darolles et al. (2004), Hall and Horowitz (2005), and Newey and Powell (2003).

Carrasco et al. (2005) and Darolles et al. (2004) consider primarily model (1.1) without an exogenous variable Z, i.e., $Y = m(X) + U$ where $\mathbb{E}(U|W) = 0$. Darolles et al. do not impose any distributional restrictions, but rather use a

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

Tikhonov regularization of the inverse of the estimated operator $A : L_2(\mu_X) \to L_2(\mu_W)$, $(A\psi)(w) = \mathbb{E}(\psi(X)|W = w)$, where they plug in kernel estimates in the expression of A. Thus, in case that joint normality holds true, their estimator is similar to the series estimator with Tikhonov regularization scheme, except that they have an additional estimation error arising from nonparametric estimation of the operator A. Carrasco et al. also consider other regularization schemes like the Landweber-Friedmann regularization. Both Carrasco et al. and Darolles et al. restrict the class of functions for m more narrowly, and then obtain polynomial rates of convergence even in the case when (X, W) are jointly normal.

Hall and Horowitz (2005) allow for a continuously distributed exogenous variable Z in model (1.1). They do not make any parametric distributional assumptions but rather assume that X, W, and Z have compact support, and therefore exclude any normally distributed components. Further, they restrict attention to the case when the problem is mildly ill-posed, meaning that the eigenvalues of the operators A_z decay at a polynomial rate, and not, as in the case of joint normality considered here, exponentially fast. Moreover, they use a density weighting that leads them to consider (in the univariate endogenous case) rates of convergence in $L_2(I)$ (I is the compact interval under consideration) instead of $L_2(\mu_X)$ (as used in Carrasco et al., 2005; Darolles et al., 2004; and BCK). Indeed, one could argue that it is more natural to use $L_2(\mu_X)$, since one should aim to estimate $m(x)$ more precisely where many observations X are available.

Newey and Powell (2003) also allow for continuously distributed Z without any additional parametric assumptions or assumptions on the support of the distribution of Z. Instead, they make stronger assumptions on the function m itself, namely that it belongs to a certain compact subset of L_2. Thus, they impose stronger (untestable) assumptions on the unknown regression function, where we prefer to restrict the distribution of (X, W, Z), which is justified in our application, and leave m as general as possible. Their estimator is a truncated series estimator in the Hermite functions (roughly the Hermite polynomials weighted by the standard normal density).

Finally, inverse problems in a statistical framework like noisy integral equations and density deconvolution problems were more generally studied within the last two decades (cf., e.g., O'Sullivan, 1986; Nychka and Cox, 1989; Donoho, 1995; Johnstone and Silverman, 1990; Mair and Ruymgaart, 1996; and Cavalier and Tsybakov, 2002).

1.3. Structure of the Paper

The paper proceeds as follows: In Section 2 we introduce the exact assumptions for model (1.1), in particular conditional normality of (X, W) given Z. In order to check this assumption in applications, we propose a novel bootstrap test for conditional normality against nonparametric alternatives. Section 2.2 contains a result on identification of m under our semiparametric specification. In contrast to fully nonparametric specifications, identification in our semiparametric framework can

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

be tested statistically. We suggest an estimator based on the principle of sample counterparts, and establish rates of convergence of this estimator, both in L_2 as well as uniformly. As an illustration, in Section 2.4 we briefly discuss the case where all variables (X, W, Z) are jointly normally distributed. In Section 3 we investigate the finite sample behavior of our estimator in a Monte Carlo experiment, before we turn to an application in Section 4. Specifically, we analyze the behavior of our estimator in a real world scenario, by using food expenditure from the British Family Expenditure Survey (FES). The greater perspective of our application is discussed in Section 5, where we argue that the conditions when a model with additive error arises are restrictive from an economic perspective. In particular, as an example we show that the the semiparametric specification of BCK is only compatible with a population with heterogeneous preferences if all individuals are identical (and the error is an orthogonal measurement error or the like), or a set of fairly implausible and untestable additional conditions hold. We conclude this paper with a summary.

2. NONPARAMETRIC IV REGRESSION WHEN INSTRUMENTS AND REGRESSORS ARE GAUSSIAN

We start with some basic assumptions and notations. Throughout we assume to have i.i.d. observations (X_i, Z_i, Y_i, W_i), $i = 1, \dots, n$ from the population model (1.1), and focus on the case where $X_i \in \mathbb{R}$ and $W_i \in \mathbb{R}$ (i.e., the univariate case) to keep the exposition concise. In Remark 2 we briefly sketch how to extend our approach to the multivariate case.

The main additional assumption that we add to the specification of model (1.1) is the following: The conditional distribution of (X, W) given $Z = z$ is normal:

$$(X, W)|Z = z \sim \mathcal{N}\left(\begin{pmatrix}\mu_1(z)\\ \mu_2(z)\end{pmatrix}, \begin{pmatrix}\sigma_1^2(z) & \rho(z)\sigma_1(z)\sigma_2(z)\\ \rho(z)\sigma_1(z)\sigma_2(z) & \sigma_2^2(z)\end{pmatrix}\right). \quad \textbf{(2.1)}$$

This assumption does not imply that the regression function $m(x, z)$ itself in (1.1) is parametrically specified, only that there are certain restrictions on the joint distribution of (X, W, Z). Moreover, we assume that $Z_i \in \mathbb{R}^d$ is continuously distributed, but this assumption can be relaxed; see the discussion after Theorem 2. This set of assumptions will be invoked without further mention.

Further, let us introduce the following notation: Let μ_{XZW} denote the joint distribution of (X, Z, W), and let the marginal distributions be denoted in a similar fashion (e.g., μ_{XW} is the distribution of (X_i, W_i)). Let f_{XZW} denote the Radon-Nikodym derivative of μ_{XZW} with respect to the Lebesgue measure (i.e., a density under our assumptions), and set $\mu_{XZ}(dx\,dz) = f_{XZ}(x, z)\,dx\,dz$, $\mu_{ZW}(dz\,dw) = f_{ZW}(z, w)\,dz\,dw$, where the marginal densities are again denoted in an obvious fashion. The conditional distribution of X given $Z = z$ is denoted by $\mu_{X|Z=z}$, with density $f_{XZ}(x, z)/f_Z(z)$, and the conditional distribution of W is similar given $Z = z$. We denote the L_2 norm with regard to the probability measure $\mu_{X|Z=z}$ by $\|\cdot\cdot\|_{\mu_{X|Z=z}}$, and similarly for $\mu_{W|Z=z}$.

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

2.1. Testing Conditional Normality

Although there is an abundance of methods to test the goodness of fit of a parametric family of distributions, tests for a parametric form of a conditional distribution are surprisingly rare. In a recent contribution, Delgado and Stute (2008) suggest tests for parametric families of conditional distributions based on the martingale transform method. Here we propose a simple test based on comparing the L_2 distance of densities using the bootstrap. We want to test whether

$$H_0: \quad \mathbb{P}(f(X,W|Z) = f(X,W|Z;\theta)) = 1$$

is true. Here, $f(x,w|z)$ denotes the joint (nonparametric) density of X, W conditional on $Z = z$, and $f(X,W|Z;\theta))$ denotes the joint density of X, W conditional on $Z = z$ under the assumption that they are normally distributed with parameter θ, depending on Z, i.e., $\theta = \theta(Z)$. The alternative is that these functions differ on a subset of the support of (X,W,Z) of positive measure. The null is equivalent to the condition that the L_2 distance of the two functions is zero. Using a nonzero and bounded weighting function a, this condition can be written as

$$\Gamma_1 = \mathbb{E}\Big(\big(f(X,W|Z) - f(X,W|Z;\theta)\big)^2 a(X,W,Z)\Big) = 0. \tag{2.2}$$

The natural sample counterpart to Γ_1 is given by

$$\widehat{\Gamma}_1 = n^{-1}\sum_i \big(\hat{f}(X_i,W_i|Z_i) - f(X_i,W_i|Z_i;\hat{\theta})\big)^2 a(X_i,W_i,Z_i), \tag{2.3}$$

where $\hat{f}(X_i,W_i|Z_i) = \hat{f}(X_i,W_i,Z_i)/\hat{f}(Z_i)$ is the ratio of two standard leave-one-out nonparametric kernel density estimators $\big($e.g., $\hat{f}(Z_i) = \sum_{j\neq i}\left(nh^d\right)^{-1} K((Z_j - Z_i)/h)\big)$, and K is a standard kernel. Moreover, $f(X_i,W_i|Z_i;\hat{\theta})$ is a semiparametric ML estimator defined as the minimizer of a local ML problem, i.e.,

$$\mathcal{L}(\theta(Z_i)) = \left(nh^d\right)^{-1}\sum_{j\neq i}\log\left[\phi\left(X_j,W_j;\theta(Z_i)\right)\right]K((Z_j - Z_i)/h),$$

where ϕ is the density of the standard normal distribution. The asymptotic distribution of the test statistic (2.3) can be obtained using arguments as in Ait-Sahalia, Bickel, and Stoker (2001), and because the large sample behavior of this test is not in the center of the paper we desist from presenting it here. Arguments for the consistency of bootstrap-based procedures for the asymptotic distribution may be found there, as well as in Härdle and Mammen (1993). The following bootstrap procedure seems natural:

1. For each $Z_i = z_i$, $i = 1,...,n$, estimate $\theta(z_i)$ in $f(X_i,W_i|Z_i = z_i;\theta(z_i))$ through semiparametric ML.

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

2. Next, draw $Z_i^* = Z_i$, from the data.
3. For each $i = 1, ..., n$, draw one observation $\left(X_i^*, W_i^*\right)$ from $f(X_i, W_i | Z_i = z_i^*; \hat{\theta}(z_i^*))$. This gives the bootstrap tuple $\left(X_i^*, W_i^*, Z_i^*\right)$.
4. From $\left(X_i^*, W_i^*, Z^*\right)$, $i = 1, ..., n$, compute $\widehat{\Gamma}_1^*$.
5. Repeat steps 2 to 4 often enough to obtain critical values for $\widehat{\Gamma}_1$.

2.2. Identification under Conditional Normality

First, to see how an integral equation arises in model (1.1), consider the conditional expectation operator

$$A : L_2(\mu_{XZ}) \to L_2(\mu_{ZW}), \qquad A\psi(z,w) = \mathbb{E}(\psi(X,Z)|Z=z, W=w) = \int \psi(x,z) \frac{f_{XZW}(x,z,w)}{f_{ZW}(z,w)}\, dx. \tag{2.4}$$

The regression function m is determined as the solution of the integral equation

$$\mathbb{E}(Y|Z=z, W=w) = (Am)(z,w). \tag{2.5}$$

Thus, both the left-hand side of (2.5) as well as the operator A are unknown in general and have to be estimated from the data.

Assume for the moment that

$$(X,W)|Z=z \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho(z)\\ \rho(z) & 1\end{pmatrix}\right), \tag{2.6}$$

and consider the conditional expectation operator given $Z = z$,

$$A_z : L_2(\mu_{X|Z=z}) \to L_2(\mu_{W|Z=z}), \qquad (A_z\phi)(w) = \int \phi(x) \frac{f_{XW|Z=z}(x,z,w)}{f_{W|Z=z}(z,w)}\, dx,$$

where obviously both $\mu_{X|Z=z}$ and $\mu_{W|Z=z}$ are the standard normal distribution for any z.

To present the spectral decomposition of A_z under assumption (2.6), consider the normalized Hermite polynomials $H_j(x)$ defined by

$$H_j(x) = (-1)^j \frac{\varphi^{(j)}(x)}{(j!)^{1/2}\varphi(x)},$$

where $\varphi(x)$ is the density of the standard normal distribution. The $(H_j)_{j\geq 0}$ form an orthonomal basis of $L_2(\mu_X)$, and

$$\left(A_z(H_k)\right)(w) = \rho^k(z)\, H_k(w), \tag{2.7}$$

i.e., they form the orthonormal bases in the singular value decomposition of A_z, where the singular values are given by $\rho^k(z)$. Now expand

$$m(x,z) = \sum_{k\geq 0} \alpha_k(z)\, H_k(x), \quad \text{where} \quad \alpha_k(z) = \int m(x,z)\, H_k(x)\, d\mu_X(x) \tag{2.8}$$

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

is the Fourier coefficient (with regard to the Hermite basis), conditional on z. Then in view of (2.7),

$$\big(A_z\, m(\cdot, z)\big)(w) = \sum_{k\geq 0} \beta_k(z)\, H_k(w) = \sum_{k\geq 0} \rho^k(z)\, \alpha_k(z)\, H_k(w),$$

where $\beta_k(z) = \mathbb{E}\big(Y H_k(W)|Z = z\big)$. Hence $\mathbb{E}\big(Y H_k(W)|Z = z\big) = \rho^k(z)\,\alpha_k(z)$, and the coefficients $\alpha_k(z)$ and hence the function m are identified on $\mathbb{R} \times I$ with $I \subset \mathbb{R}^d$ if and only if $\rho(z) \neq 0$ for all $z \in I$ (or at least μ_Z-almost all $z \in I$). In the general case (2.1), since all the objects μ_i and σ_i^2, $i = 1, 2$, are identified, we can simply assume that $(X, W)|Z$ is standardized. Thus, we obtain a theorem giving sufficient and necessary conditions for identification under conditional normality.

THEOREM 1. *Under assumption (2.1), the function $m(x, z)$ in model (1.1) is identified on $\mathbb{R} \times I$ with $I \subset \mathbb{R}^d$ if and only if $\rho(z) \neq 0$ for μ_Z-almost all $z \in I$.*

Remark 1. In the absence of restrictions on the joint distribution of (X, W, Z), in order to achieve identification one has to require that the conditional operators A_z are one-to-one for (almost every) z or, equivalently, that the singular values $\lambda_{k,z}$, $k \geq 1$, are nonzero for each z. This is clearly not a testable assumption, since it involves all the $\lambda_{k,z}$ simultaneously. In contrast, under the (testable) assumption (2.1), one can establish identification. Suppose that $\hat{\rho}(z)$ is a nonparametric estimate of $\rho(z)$, that does not change sign (otherwise identifiability is questionable), without loss of generality assume that $\hat{\rho}(z) > 0$. For a compact interval I with a high concentration of observations Z_i, one can now bootstrap the distribution of $\inf_{z\in I} \hat{\rho}(z)$. If the 5% quantile of this distribution is larger than zero, the hypothesis of identification is not rejected.

Remark 2. The diagonalization of the conditional expectation operators A_z in terms of the Hermite polynomials is based on the so-called Mehler's formula, for which also multivariate and, in particular, bivariate extensions exist (cf. Erdelyi, 1939).

2.3. Estimation under Conditional Normality

Now let us turn to estimation under Assumption 2.5. Let $\hat{\mu}_i$, $\hat{\sigma}_i$, $i = 1, 2$, and $\hat{\rho}(z)$ be nonparametric estimates of μ_i, σ_i, and $\rho(z)$ respectively. Generally, we have $\mathbb{E}\big(Y H_k(W^*)|Z = z\big) = \beta_k(z)$, where $W^* = (W - \mu_2(Z))/\sigma_2(Z)$. Therefore, using the estimated normalized variables

$$\hat{W}_i^* = \frac{W_i - \hat{\mu}_2(Z_i)}{\hat{\sigma}_2(Z_i)},$$

$\beta_k(z)$ can be estimated by nonparametric regression. We shall use a local linear estimator and for simplicity, restrict the following discussion to the case $d = 1$.

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

Then the estimator for $\beta_k(z)$ is given by

$$\hat{\beta}_k(z) = \frac{\sum_{j=1}^n l_j(z) Y_j H_k(\hat{W}_j^*)}{\sum_{j=1}^n l_j(z)}, \tag{2.9}$$

where

$$l_j(z) = K\Big(\frac{z - Z_j}{h}\Big)\big(s_{n,2}(z) - (z - Z_j)s_{n,1}(z)\big),$$
$$s_{n,i}(z) = \sum_{j=1}^n K\Big(\frac{z - Z_j}{h}\Big)(z - Z_j)^i, \qquad i = 1, 2,$$

K is a kernel function, and $h > 0$ is the bandwidth.

The final estimator of $m(x, z)$ is hence given by

$$\hat{m}(x, z) = \sum_{k \geq 0} \hat{\beta}_k(z)\, w\big(\hat{\rho}(z), k; M\big)\, H_k\Big(\frac{x - \hat{\mu}_1(z)}{\hat{\sigma}_1(z)}\Big). \tag{2.10}$$

Here, $w(\cdot, \cdot)$ is a regularization scheme, and M its regularization parameter. For example, if $\hat{\rho}(z) \neq 0$ for all z, one can use a simple truncation scheme (spectral cutoff)

$$w_{sco}\big(\hat{\rho}(z), k; M\big) = \begin{cases} \hat{\rho}(z)^{-k}, & k \leq M, \\ 0 & k > M. \end{cases}$$

Alternatively, the Tikhonov regularization scheme with regularization parameter $\alpha > 0$ ($\alpha \to 0$ as $n \to \infty$) is given by

$$w_{Tyk}(\hat{\rho}(z), k; \alpha) = \frac{\hat{\rho}(z)^k}{\alpha + \hat{\rho}(z)^{2k}}.$$

We analyze the asymptotic behavior of this estimator under the set of assumptions below. Here, $C > 0$ is an arbitrarily large but fixed constant.

Assumption 1. The coefficients $\alpha_k(z)$ in the expansion (2.8) satisfy $|\alpha_k(z)| \leq Ck^{-\gamma}$, $\gamma > 3/4$, uniformly in z, and are two times continuously partially differentiable with uniformly bounded (in z and k) first and second derivatives.

Assumption 2. The functions $\mu_i(z)$ and $\sigma_i(z)$, $i = 1, 2$, and $\hat{\rho}(z)$ are two times continuously partially differentiable and uniformly bounded up to the second derivative by $C > 0$. Further, $1 > \rho_{\max} \geq |\rho(z)| \geq \rho_{\min} > 0$ for $z \in I \subset \mathbb{R}$, and m is uniformly bounded by $C > 0$.

Assumption 3. The estimators $\hat{\mu}_i(z)$, $\hat{\sigma}_i(z)$ and $\hat{\rho}(z)$ for $\rho(z)$ converge uniformly on compact intervals $I \subset \mathbb{R}$ with a polynomial rate; e.g., there is a $\epsilon_0 > 0$ such that $\sup_{z \in I} |\hat{\rho}(z) - \rho(z)| = O(n^{-\epsilon_0})$.

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

Assumption 4. The density f_Z of the Z_j is bounded away from 0 by $1/C$ on compact intervals $I \subset \mathbb{R}$.

Assumption 5. The kernel function K in the local linear estimators (2.9) is a bounded, symmetric probability density function, $\int zK(z)\,dz = 0$, $\int z^2K(z)\,dz < \infty$, and $\int K^2(x)\,dx < \infty$. The bandwidth is chosen as $h \sim n^{-1/5}$.

Assumptions 1 and 2 are similar to assumption MV.3 in Hall and Horowitz (2005). Assumptions 4 and 5 are standard in nonparametric regression and are taken from Fan (1992). Finally, note that in Assumption 3 we do not further specify the nonparametric estimators of the functions $\mu_i(z)$, $\sigma_i(z)$, and $\rho(z)$. Uniform convergence is proved for $\mu_i(z)$ in Blondin (2007) for the local linear estimator, and for $\sigma_i(z)$ and $\rho(z)$ in Neumann (1994) for kernel estimators.

The result below, whose proof can be found in the Appendix, summarizes the asymptotic behavior.

THEOREM 2. *Suppose that $I \subset \mathbb{R}$ is compact. Let Assumptions 1–5 be true. For the estimator (2.10) based on the spectral cutoff regularization scheme, for $M = c\log n$ with sufficiently small c we have that for every $\delta > 0$ there is C_δ and n_0 such that for all $n \geq n_0$, and for all $z \in I$,*

$$\mathbb{P}\Big(\|\hat{m}_{sco}(\cdot,z) - m(\cdot,z)\|^2_{L_2(\mu_{X|Z=z})} > C_\delta\big(\log n\big)^{-2\gamma+1}\Big) < \delta. \tag{2.11}$$

Further, for compact $J \subset \mathbb{R}$ we have that for all $z \in I$,

$$\mathbb{P}\Big(\sup_{x\in J}|\hat{m}_{sco}(x,z) - m(x,z)| > C_\delta\big(\log n\big)^{-\gamma+3/4}\Big) < \delta. \tag{2.12}$$

Remark 3. The first part of the theorem, equation (2.11), is similar to Theorem 4.3 in Hall and Horowitz (2005), though they consider operators A_z, which are mildly ill-posed. The rate of convergence obtained in their setting also depends on the additional exogenous regressors Z, which is surprisingly not the case for the severely ill-posed problem with conditional normal endogenous regressors and instruments as considered in Theorem 2 (see the proof of Theorem 2 in the Appendix for the mathematical reason).

2.4. Identification and Estimation under Full Joint Normality

In general the eigenvalues $\rho^k(z)$ of the operator A_z in Section 2.2 will depend on z, and although the rate is independent of the magnitude of $\rho(z)$, the relevant constants will strongly depend on it. Therefore, we investigate the dependence of $\rho(z)$ on z in a particular case, namely when we have joint normality of the vector (X, W, Z). It turns out that in this important special case, $\rho(z)$ will be constant. We shall suggest an alternative estimator under full joint normality and study its large sample behavior.

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

Assume for simplicity that the (X, Z, W) have been standardized, so that they are jointly normally distributed with mean-vector zero and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho_{XZ} & \rho_{XW} \\ \rho_{XZ} & 1 & \rho_{ZW} \\ \rho_{XW} & \rho_{ZW} & 1 \end{pmatrix}. \tag{2.13}$$

Then the entries in the conditional distribution (2.1) are given by $\mu_1(z) = \rho_{XZ} z$, $\mu_2(z) = \rho_{WZ} z$, and $\sigma_1^2 = 1 - \rho_{XZ}^2$, $\sigma_2^2 = 1 - \rho_{WZ}^2$, and

$$\rho = \big(\rho_{XW} - \rho_{XZ}\rho_{ZW}\big)\big((1-\rho_{XZ}^2)(1-\rho_{WZ}^2)\big)^{-1/2}, \tag{2.14}$$

so that σ_1, σ_2, and ρ are in this case independent of z. Thus, global regularization in z is reasonable in this case. Also note that identification of m is now equivalent to $\rho \neq 0$, which can easily be tested parametrically. Under joint normality, one can also give the spectral decomposition of the unconditional operator A. It is in fact easy to show that the functions

$$\psi_{j,k}(x,z) = H_j\big(\frac{x - \rho_{XZ} z}{\sigma_1}\big) H_k(z), \qquad \chi_{j,k}(z,w) = H_j\big(\frac{w - \rho_{ZW} z}{\sigma_2}\big) H_k(z),$$

form orthonormal bases of the spaces $L_2(\mu_{XZ})$ and $L_2(\mu_{ZW})$, respectively, and

$$(A\psi_{j,k})(z,w) = \rho^j \chi_{j,k}(z,w),$$

where ρ is given in (2.14). Therefore, under joint normality one can also use a double series estimator. Let

$$\beta_{j,k} = \int (Am)(z,w)\, \chi_{j,k}(z,w)\, \mu_{ZW}(dz\, dw),$$

$$\alpha_{j,k} = \int m(x,w)\, \psi_{j,k}(x,w)\, \mu_{XW}(dx\, dw).$$

An estimator with a truncation regularization scheme is given as

$$\hat{m}_{JN}(x,z) = \sum_{j=0}^{M_1} \sum_{k=0}^{M_2} \frac{\hat{\beta}_{j,k}}{\hat{\rho}^j}\, \hat{\psi}_{j,k}(x,z), \tag{2.15}$$

where M_1 and M_2 are truncation parameters, and

$$\hat{\beta}_{j,k} = \frac{1}{n} \sum_{i=1}^{n} Y_i H_j\big(\frac{W_i - \hat{\rho}_{WZ} Z_i}{\hat{\sigma}_2}\big) H_k(Z_i), \qquad \psi_{j,k}(x,z) = H_j\big(\frac{x - \hat{\rho}_{XZ} z}{\hat{\sigma}_1}\big) H_k(z).$$

Further, one can, e.g., choose $\hat{\rho}_{XZ} = \frac{1}{n}\sum_{k=1}^{n} X_k Z_k$, and $\hat{\rho}_{ZW}$ and $\hat{\rho}_{XW}$ are similar parametric estimators. Further, $\hat{\rho}$, $\hat{\sigma}_1$, and $\hat{\sigma}_2$ are obtained by plugging them into the corresponding expressions for ρ, σ_1, and σ_2, respectively. The convergence rates (2.11) and (2.12) also hold for this estimator. Moreover, one can give an even more global convergence result (with regard to the joint distribution μ_{XW}) under the assumptions below.

Assumption 6. The coefficients $\alpha_{j,k}$ of m satisfy $|\alpha_{j,k}| \leq C j^{-\gamma} k^{-\delta}$ for some $\gamma, \delta > 1/2$.

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

Assumption 7. The conditional variance function $\sigma^2(w,z) = \mathbb{E}(U^2|Z = z, W = w)$ and m are uniformly bounded (bounded in $\|\cdot\|_\infty$).

Assumption 8. The parameter estimates $\hat{\rho}_{XZ}$, $\hat{\rho}_{ZW}$, $\hat{\rho}_{XW}$, $\hat{\rho}$, $\hat{\sigma}_1$, and $\hat{\sigma}_2$ converge in probability with parametric rate $n^{-1/2}$.

Assumption 6 is a standard smoothness assumption for the function m, and Assumption 7 is also rather common in nonparametric regression. Assumption 8 is satisfied, e.g., by maximum likelihood estimators.

THEOREM 3. *Suppose that (X, W, Z) are jointly normal and standardized with covariance structure given in (2.13), and that Assumptions 6, 7, and 8 are satisfied. Then for $M_1 = c \log n$ with $c > 0$ small enough and $M_2 = n^{-\epsilon}$ for $\epsilon > 0$ small, we have for the estimator $\hat{m}_{JN}$ in (2.15) that*

$$\|\hat{m}_{JN} - m\|^2_{L_2(\mu_{XZ})} = O_P\left((\log n)^{-2\gamma+1}\right). \tag{2.16}$$

This appears to be the first result in which a rate of convergence is derived for a nonparametric IV estimator with additional continuous exogenous variables in $L_2(\mu_{XZ})$, i.e. with regard to the joint distribution of (X, Z). Again, the rate in (2.16) does not depend on the additional regression on Z.

Remark 4. If exogenous regressors Z_i are present in the regression problem (1.1), then the unconditional operator A in (2.4) is no longer a Hilbert-Schmidt operator, since it acts as the identity on functions $g(z) \in L_2(\mu_Z) \subset L_2(\mu_{XZ})$. Therefore, the image A^*Am can no longer be estimated at a fast rate (in case of known A that is Hilbert-Schmidt, this rate is $n^{-1/2}$), and the strategy in Darolles et al. (2006) to first estimate A^*Am and then to apply a regularized version of the inverse of A^*A (or its estimated version) can no longer be pursued. In Section 2.3 (and similarly in Hall and Horowitz, 2005), this is overcome by assuming that the conditional operators A_z are Hilbert-Schmidt and moreover satisfy a uniformity condition (expressed in the assumption that $\rho(z)$ is bounded away from 0). In this section under joint normality, we show that one can simply construct a bivariate series estimator that covers both the identity component of A as well as the Hilbert-Schmidt component, and for which two distinct truncation parameters are required that grow at different rates.

3. SIMULATION

In this section we report the results of an extensive simulation study. We start with a scenario with conditional normality, and generate data (Y_i, X_i, Z_i, W_i) from model (1.1) for distinct functions m_j, where

$$m_1(x,z) = (-1.5x)(z+2), \qquad m_2(x,z) = (x^2+2)(z^2x+0.5),$$
$$m_3(x,z) = \exp(-0.2(z+2)x).$$

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

For the (X, Z, W, U), generate a three-dimensional normal random vector (U, X^*, W^*) with mean-vector zero and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.7 \\ 0 & 0.7 & 1 \end{pmatrix}.$$

Further, draw Z independent of (U, X^*, W^*) from a beta distribution $B(1/2, 1/3)$, and standardize it. Then set $X = X^* + Z^2$ and $W = W^* - (Z^3 - Z)/2$. Thus, we have $\mu_1(z) = z^2$, $\mu_2(z) = -(z^3 - z)$, $\sigma_1^2(z) = \sigma_2^2(z) = 1$, and $\rho(z) = 0.7$ are constant. We estimate these functions nonparametrically by a local polynomial estimator of order one with a data-driven plug-in bandwidth choice. Then, we simulate 10,000 times samples of size $N = 5,000$ for each m_j, and use the estimator (2.10) with the simple truncation scheme and different truncation parameters. As in (2.11), we are mainly interested in the behavior of the m_j as a function of x for fixed z. Therefore, for two distinct values of z ($z_1 = 0.1$, $z_2 = 1$) we record the values of the coefficient $\hat{\beta}_k(z_i)$, $k = 0, \ldots, 4$ and $i = 1, 2$. Further, for each of these truncation parameters we compute $\|\hat{m}(x, z_i) - m_j(x, z_i)\|^2_{L_2(X|Z=z_i)}$ (the weighted error) and $\|\hat{m}(x, z_i) - m_j(x, z_i)\|^2$ (the unweighted error over the interval $[-2.5, 2.5]$).

The results for estimation at z_1 are presented in Tables 1–3; for z_2, the results are similar though a little less good. We see that for suitably chosen truncation parameter M, the estimator performs well for all target functions m_j, $j = 1, 2, 3$. However, if M is chosen too high, then the estimation precision decreases significantly, in particular the 0.75 and 0.95 quantiles of the distribution of the (weighted

TABLE 1. Results for estimation of m_1 at z_1

Spectral cutoff regularization					
Truncation parameter		$M = 0$	$M = 1$	$M = 2$	$M = 3$
Weighted MISE	median	9.67	0.01	0.04	0.21
	upper quartile	9.70	0.01	0.11	0.51
	0.95 quantile	9.76	0.07	0.43	1.85
Unweighted MISE	median	109.70	0.04	0.56	2.41
	upper quartile	109.82	0.11	1.55	5.59
	0.95 quantile	110.15	0.70	6.04	19.51
Coefficient		$\beta_0(z_1)$	$\beta_1(z_1)$	$\beta_2(z_1)$	$\beta_3(z_1)$
	median	−0.08	−3.16	−0.02	0.09
	IQR	0.18	0.09	0.36	0.80
	0.95-quantile	0.14	−3.02	0.48	1.23
	0.05-quantile	−0.31	−3.31	−0.51	−1.05

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

TABLE 2. Results for estimation of m_2 at z_1

Spectral cutoff regularization					
Truncation parameter		$M=0$	$M=1$	$M=2$	$M=3$
Weighted MISE	median	0.45	0.46	0.06	0.14
	upper quartile	0.45	0.49	0.12	0.31
	0.95 quantile	0.47	0.55	0.38	1.07
Unweighted MISE	median	6.54	6.87	0.68	1.84
	upper quartile	6.85	7.24	1.54	4.14
	0.95 quantile	7.33	7.91	5.48	13.25
Coefficient		$\beta_0(z_1)$	$\beta_1(z_1)$	$\beta_2(z_1)$	$\beta_3(z_1)$
	median	1.50	0.19	0.79	0.18
	IQR	0.10	0.17	0.31	0.57
	0.95-quantile	1.63	0.38	1.31	1.08
	0.05-quantile	1.38	−0.03	0.41	−0.50

TABLE 3. Results for estimation of m_3 at z_1

Spectral cutoff regularization					
Truncation parameter		$M=0$	$M=1$	$M=2$	$M=3$
Weighted MISE	median	0.22	0.02	0.01	0.02
	upper quartile	0.22	0.03	0.02	0.05
	0.95 quantile	0.22	0.03	0.06	0.19
Unweighted MISE	median	2.71	0.29	0.12	0.28
	upper quartile	2.74	0.34	0.26	0.67
	0.95 quantile	2.80	0.44	0.86	2.16
Coefficient		$\beta_0(z_1)$	$\beta_1(z_1)$	$\beta_2(z_1)$	$\beta_3(z_1)$
	median	1.07	−0.43	0.12	−0.01
	IQR	0.05	0.06	0.13	0.24
	0.95-quantile	1.13	−0.36	0.30	0.31
	0.05-quantile	1.02	−0.52	−0.08	−0.38

or unweighted) mean integer square error (MISE) increase. This effect is stronger for the unweighted case then for the weighted case. A similar effect can be observed for the coefficients $\beta_j(z_i)$. The estimates of higher-order coefficients have high variance. Thus, even though the median of these estimates is reasonable, the variability becomes large, which can lead to very bad estimators in some

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

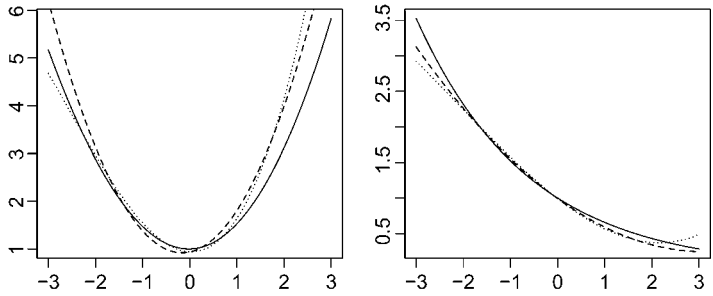


FIGURE 2. Left: True target function m_2 (solid line) and estimates with spectral cutoff regularization scheme with $M = 2$ (dashed line) and $M = 3$ (dotted line). Right: True target function m_3 (solid line) and estimates with spectral cutoff regularization scheme with $M = 2$ (dashed line) and $M = 3$ (dotted line).

cases if these coefficients are included. Nevertheless, since only a few (small and discrete) values of M have to be tested, a proper choice of the smoothing parameter is no problem in practice. See also Figure 2 for plots of the true functions and typical estimates. As a benchmark, we also performed simulations for estimating m_2 at z_1, but where the exogenous covariate is fixed at z_1. Although there is no asymptotic effect when regressing on Z, there is quite a finite sample effect; compare Tables 4 and 2. One reason is the automatic bandwidth choice in the nonparametric estimation of the various quantities by a local linear

TABLE 4. Results for estimation of m_2 at z_1, where z_1 is fixed, i.e., without exogenous component

Spectral cutoff regularization					
Truncation parameter		$M = 0$	$M = 1$	$M = 2$	$M = 3$
Weighted MISE	median	0.44	0.44	0.01	0.02
	upper quartile	0.44	0.44	0.01	0.03
	0.95 quantile	0.44	0.44	0.03	0.07
Unweighted MISE	median	6.62	6.61	0.08	0.20
	upper quartile	6.70	6.69	0.17	0.39
	0.95 quantile	6.80	6.81	0.39	0.85
Coefficient		$\beta_0(z_1)$	$\beta_1(z_1)$	$\beta_2(z_1)$	$\beta_3(z_1)$
	median	1.49	0.05	0.69	0.02
	IQR	0.02	0.07	0.11	0.20
	0.95-quantile	1.52	0.13	0.82	0.27
	0.05-quantile	1.46	−0.03	0.56	−0.10

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

estimator in the setting with exogenous covariate. We use a direct plug-in rule for estimating the bandwidth, which performs reasonably well in the majority of cases, but sometimes is also much too small, so that the estimator behaves rather poorly.

We furthermore conducted simulations with a nonconstant correlation function, in order to investigate its effect, in particular concerning the uniform choice of the truncation parameter M. We chose $\rho(z) = 0.6 + z\,0.3$, where $Z \sim B(1.5, 1.5)$, and considered the regression function $m(x, z) = x^2 + 2$, which does not depend on z (so that the same conditional function is estimated). We chose $z_1 = 0.2$ and $z_2 = -0.2$, and compared the estimation results. However, these did not differ substantially; the errors when estimating nonparametrically the functions $\beta_k(z)$ and $\rho(z)$ by local linear estimation had the most effect. Further, since only small values for M are required and $\rho(z)$ is comparatively large, the error when dividing by the small power of $\hat{\rho}(z_i)$ does not have much more effect than the error when estimating $\hat{\beta}_k(z_i)$.

Finally, in a different scenario under full joint normality of (X, W, Z), we compare the performance of the estimators (2.10) (which only uses conditional normality) and (2.15) (which exploits the full joint normality). In turns out that the estimator (2.15) performs only slightly better than the estimator (2.10), even though it takes stronger advantage of the distributional properties. Thus, in practice we recommend choosing the estimator (2.10), which has less stringent distributional assumptions.

4. APPLICATION

4.1. Empirical Specification

Neglecting the issue of preference heterogeneity for the moment, following BCK we simply assume that there is a relationship m between a K-vector of dependent variables Y, in our case budget shares, and total expenditure X as well as a set of exogenous regressors Z such that

$$Y = m(X, Z) + U, \qquad \mathbb{E}[U|W, Z] = 0 \tag{4.1}$$

holds, where W is disposable income. The reason that X is believed to be endogenous in demand analysis is that total expenditure and demand for individual goods are both believed to be parts of a general preference ordering of individuals, which is partly incorporated in the error U. Disposable income of individuals, in contrast, is believed to be exogenous because the driving unobservables like ability are assumed to be independent of the preference orderings. Of course, this assumption is questionable, but in this paper we follow the demand literature in assuming that this is the case. Additionally, we assume that preferences are such that they admit a sufficiently smooth demand relationship (e.g., preferences are strictly convex, with associated differentiable utility functions). We use household data, but abstract from any interfamily allocation issues, which we feel is justified given the focus of this paper.

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

At the core of the paper is the joint normality of X and W, conditional on Z, which we argue is a reasonable assumption in our application. Indeed, our approach aims at exploiting this knowledge about the distribution, or turned a bit more negatively, argues that even using this knowledge the situation is difficult. One advantage associated with this semiparametric element is the standard one: It allows us to mitigate the curse of dimensionality. This is, however, not the only feature we employ to make nonparametric analysis tractable in the high-dimensional environment that we are working in. Specifically, when conditioning on household observables, we use a low-dimensional vector of principal components instead of the original high-dimensional vector of household covariables. Like much of the demand literature, we implicitly condition on some of them because we are working in a specific subpopulation, defined by household characteristics. However, the original data have more than 100 household covariates, many of which are discrete. Including them in an unrestricted nonparametric analysis is impossible. Instead, we have performed some robustness checks for our use of principal components. In a mild abuse of notation we refer to these components from now on also as Z. We now turn to a precise description of the data.

4.2. Data Description

For our analysis we employ the British FES. Every year, the FES reports the income, expenditures, demographic composition, and other characteristics of about 7,000 households. The sample surveyed represents about 0.05% of all households in the United Kingdom. The information is collected partly by interview and partly by records. Records are kept by each household member, and include an itemized list of expenditures during 14 consecutive days. The periods of data collection are evenly spread out over the year. The information is then compiled and provides a repeated series of yearly cross sections.

The category of goods we consider is food related, and consists of the subcategories food bought and catering, which are self-explanatory. Together our food category accounts for 28% of expenditures on average. We remove outliers by excluding the upper and lower 2.5% of the population. Income is constructed as in the definition of the "household below average income study" (HBAI). It is roughly defined as net income after taxes, but including state transfers. This is done in both data sets to define nominal income. Real income is then obtained by dividing through the retail price indexes. As mentioned, we stratify the population to obtain a relatively homogeneous subpopulation, which is equivalent to controlling for the influence of discrete controls nonparametrically. Like much of the demand literature, we focus on one subpopulation to minimize measurement error.

4.3. Results

We first start by reporting the results for the test of conditional normality. The value of the test statistic $\widehat{\Gamma}_1$ is 0.13463, which corresponds to a p-value of the

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

bootstrap distribution of 0.85. This indicates clearly that conditional normality is not rejected. The stronger hypothesis of joint normality, however, is rejected using a standard normality test with a p-value of virtually zero.

Since our analysis is nonparametric, many of our findings can be conveniently summarized in univariate graphs. We will be particularly concerned by the following issues, which will be illustrated by a comparative graph: Sensitivity of our results with respect to the choice of the truncation parameter M, and comparison with the exogenous and the control function IV case.

Start with the choice of M, the degree of the polynomial. Figure 3 illustrates the effect of variations in M on the conditional normal estimator (CNE). While the results remain relatively stable when moving from $M = 1$ to $M = 3$, with only $M = 4$ we obtain implausible economic structures like the upswing of the food budget share at higher levels of total expenditure. Indeed, for $M > 4$ this upswing increases further, rendering this method ineffective. This is in accordance with the simulation results, which show that the variance of the estimator increases significantly with increasing M; see Section 4.

The figure shows the behavior of the estimator at the center of Z distribution. The results remain unchanged, with slight variations in Z, but for larger deviations from the center of the distribution we again obtain economically implausible results. Here it has to be taken into account that the estimator is designed for

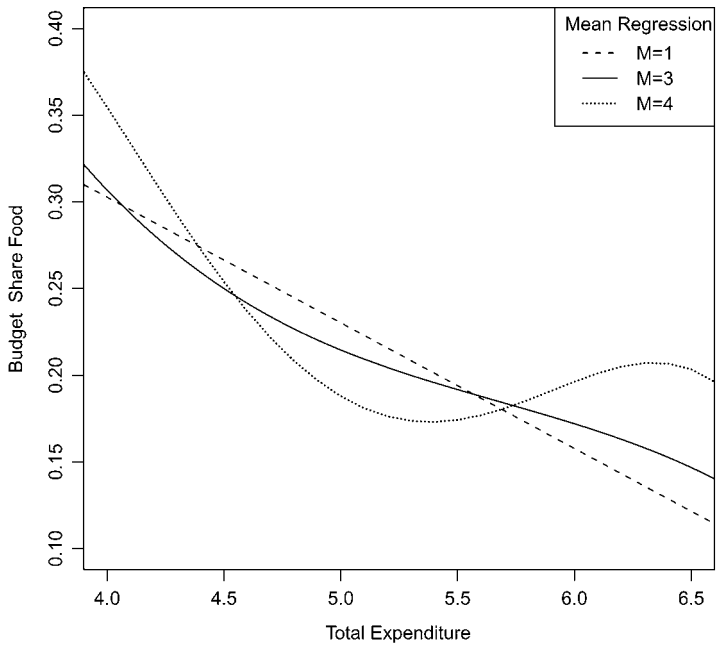


FIGURE 3. CN-estimator with varying choice of the truncation parameter M ($M = 1, 3, 4$).

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

estimation in $L_2(\mu_{X|Z=z})$, and thus cannot be expected to be very precise for values of Z that are far from the center. Moreover, this suggests that different values of M, the degree of Hermite polynomial, are required for different values of Z.

The effect of correcting for endogeneity is analyzed in Figure 4, which shows a comparison of an exogenous nonparametric regression with our CNE. Moreover, an alternative way of controlling for endogeneity is also displayed: control function IV. Indeed, we find only moderate evidence that correcting for endogeneity does affect the results. Certainly, the downward trend in food budget shares is unambiguous and very pronounced regardless of the specifications. If anything, correcting for endogeneity does seem to have a mitigating effect on the marginal effect. This is most pronounced when comparing the exogenous regression with control function IV, while our CNE is somewhere in the middle.

These findings seem to suggest that our estimators are robust to misspecification. However, they are not: If instead of conditional normality we employ (erroneously) joint normality, much less sensible results appear. We conclude that the CNE is sensitive to changes in the specification. That this effect is not apparent when analyzing endogeneity in Figure 4 is more of an indication that the effect of endogeneity is rather moderate in our application (which is in surprising contrast to BCK, who used very similar data). How the estimator performs in a setting where the regressors are strongly endogenous remains to be determined.

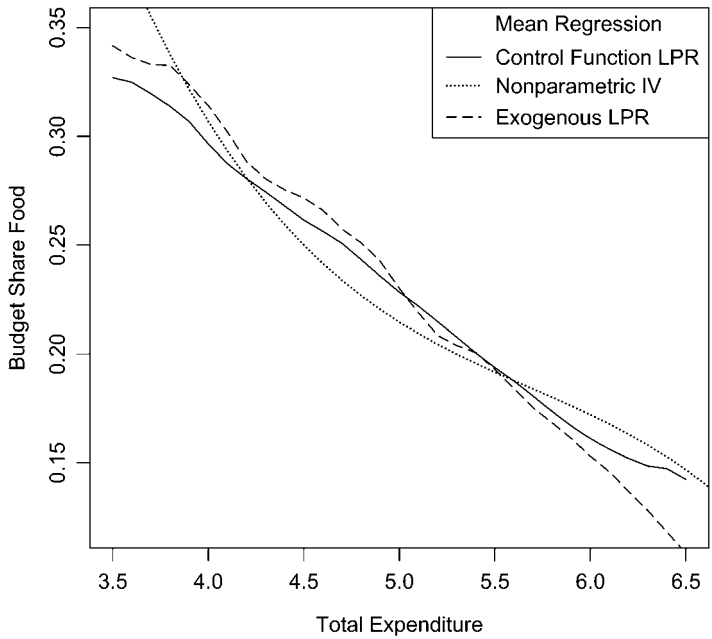


FIGURE 4. Effect of correcting for endogeneity.

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

5. ARE NONPARAMETRIC IV AND HETEROGENEOUS POPULATIONS COMPATIBLE IN EMPIRICAL WORK?

In the previous sections we have shown that feasible nonparametric IV techniques can be proposed and implemented even in situations where the problem at hand is severely ill-posed. Since many continuous variables in economic applications are approximately normally distributed, our results are encouraging for the applicability of this method. However, at this point we would like to raise an issue that suggests that we apply these types of methods with caution and only after careful deliberation, at least in some economic applications. This issue is unobserved heterogeneity, and the related question about the identification of structural parameters in a heterogeneous population. Indeed, we will argue in this section that the mean independence restriction we employ, namely $\mathbb{E}[U|W,Z]=0$, is not likely to identify meaningful parameters of interest in a heterogeneous population.

We illustrate this issue with the model of BCK. We take their approach, because their paper is (to the best of our knowledge) the only successful attempt for application of the techniques put forward in the econometric literature on ill-posed inverse problems, and hence is most closely related to our work. Also, the point we make can be illustrated best in the semiparametric specification they suggest. But it is by no means confined to their paper, and hence is not a specific drawback of their approach. Indeed, it will become obvious that if anything this point aggravates in the entirely nonparametric approach. Moreover, the approach put forward in our paper does not really offer a means to circumvent this problem; it (only) allows us to deal with observable heterogeneity in a slightly more general fashion.

But let us sketch the problem: *On the individual level,* BCK consider the class of preferences giving rise to following model:

$$Y = h(X - Z'\theta_1) + Z'\theta_2, \tag{5.1}$$

where, in their semiparametric specification, $\theta = \left(\theta_1', \theta_2'\right)'$ denotes a (individual specific) vector of coefficients, and h an unrestricted smooth function[2]. *On the individual level*, this model can be shown to be in line with the economic theory; see Pendakur (1999). Of course, as in much of microeconometrics, BCK estimate their model with data coming from a heterogeneous population. BCK acknowledge this fact by appending an additive error U, and assuming that $\mathbb{E}[U|W,Z]=0$ holds, motivating the need for an IV restriction by the correlation of total expenditure contained in X and unobserved preference heterogeneity contained in U.

The question one might ask is why does an appended additive error capture preference heterogeneity at all? Starting with Brown and Walker (1989), the demand literature has questioned the use of an additive error as a means of capturing unobserved heterogeneity. Indeed, when moving from the individual level model (5.1) to a heterogeneous population, it would be quite natural to consider

$$Y_i = h_i(X_i - Z_i'\theta_{1i}) + Z_i'\theta_{2i}, \qquad i = 1, \ldots, n. \tag{5.2}$$

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

Obviously, this model is distinct from the one considered in BCK. If the population were completely identical, i.e., $\forall i : h_i = h, \theta_i = \theta$, these common coefficients h, θ could be estimated from the data. But in this case there were also no residual U and no endogenous regressor. Returning to the unrestricted model defined through equation (5.2), in Hoderlein, Klemelä, and Mammen (2007), the estimation of the distribution of coefficients f_θ in the much simpler model of the form $Y_i = Z_i'\theta_i$ is discussed, but model (5.2) is beyond the scope of their approach as it stands.

However, a distribution of marginal effects is not what is of interest in the nonparametric IV literature. Hence, in this section and throughout the rest of the paper, we assume that interest centers on some average or "typical" effect in a heterogeneous population. Since BCK are not explicit about this issue, we assume it is the mean value of the coefficient (denoted $\overline{\theta}$) across the heterogeneous population.[3] From now on, for the rest of this section we will drop the subscript i on the coefficients, and assume that all variables other than the means $\overline{\theta}$ vary across the population. To make life simple, we assume that the model on the individual level is the less general

$$Y = (X - Z'\theta_1)\theta_3 + Z'\theta_2,$$

where θ_3 is a scalar random coefficient; i.e., $h(\zeta) = \theta_3\zeta$. In loose analogy to the semiparametric efficiency literature, we could argue that the case with general h cannot be better than the worst parametric submodel. It is easy to show that we may rewrite this model as

$$Y = (X - Z'\overline{\theta}_1)\overline{\theta}_3 + Z'\overline{\theta}_2 + U,$$

with $U = \left(\theta_3 - \overline{\theta}_3\right) X + Z' \left\{ \left(\theta_1 - \overline{\theta}_1\right)\overline{\theta}_3 + \left(\theta_3 - \overline{\theta}_3\right)\theta_1 + \left(\theta_2 - \overline{\theta}_2\right) \right\}$. The remaining question is now, How do we arrive at $\mathbb{E}[U|W, Z] = 0$?

To this end, we assume that we have instruments that are independent of unobserved preference heterogeneity. In our setup, assume that $(W, Z) \perp \theta$, which corresponds also to the economic motivation BCK provide. Then,

$$\begin{aligned}\mathbb{E}[U|W, Z] &= \mathbb{E}\left[\left(\theta_3 - \overline{\theta}_3\right) X|W, Z\right] + \mathbb{E}\left[\left(\theta_3 - \overline{\theta}_3\right) Z'\theta_1|W, Z\right] \\ &= \mathrm{Cov}\left\{\theta_3, V + Z'\theta_1|W, Z\right\},\end{aligned}$$

where $V = X - \mathbb{E}[X|W, Z]$. Even full joint independence of (θ, V) from (W, Z) is not sufficient to reduce $\mathbb{E}[U|W, Z]$ to a constant, let alone to zero. Consequently, we have to *assume* that $\mathrm{Cov}\left\{\theta_3, V + Z'\theta_1|W, Z\right\} = 0$ holds, for which there is no plausible economic reason. Of course, for any other parametric specification of h, a similarly implausible untestable restriction would be required to hold, and for the model with unknown infinite dimensional parameter h it would be the envelope of these restrictions. This effectively rules out that a model like $Y = h(X - Z'\theta_1) + Z'\theta_2$ holds on the individual level with meaningful heterogeneity in parameters across the population. If we believe the model with errors

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

mean independent of instruments to be correct, then we have effectively assumed that the population shares the identical structural model defined by m, up to an additive shift parameter that varies and is endogenous. Needless to mention, this is quite restrictive in some applications.

This type of observation has predecessors: Other than Brown and Walker (1989), Lewbel (2001) has made a related observation for a heterogeneous "almost ideal" population in the purely exogenous setting. Moreover, in the literature in economic theory on aggregation, findings that are similar in spirit emerge; see Hildenbrand (1993). The deeper reason is that models defined by some type of conditional mean restriction implicitly aggregate across a heterogeneous population, and aggregation is a step by which many properties vanish. Instead, features of the aggregation process, in our example conditional covariances, are introduced. And the nonparametric regression model involving an additive error U and obeying a restriction $\mathbb{E}[U|W, Z] = 0$ is just a case in point.

What are the alternatives? If the goal is to get an overview of the distribution of the coefficients, one possible route is the afforementioned random coefficients model of Hoderlein et al. (2007). This framework can allow for preference heterogeneity and regressors that are endogenous. However, the drawback is that it is, as of yet, limited to linear models on an individual level. Another alternative is nonseparable models as in Hoderlein (2008), Hoderlein and Mammen (2007), or Imbens and Newey (2007), which can allow for endogenous regressors even in setups that are more general than the one considered by BCK.

The drawback of these alternative approaches is that they employ control function methods to correct for endogeneity, i.e., they require that the control functions V be correctly specified and hence require correct specification of the relationship that defines it, e.g., $V = X - \mathbb{E}[X|W, Z]$. It is the advantage of the ill-posed inverse literature that this equation need not be specified, and no risk of misspecification arises. But, as we have seen, there may be a price to pay in terms of economic foundation in a heterogeneous population.

6. CONCLUSION AND OUTLOOK

This paper is concerned with analyzing issues in the application of nonparametric IV regression models, which lead to ill-posed inverse problems for economic questions. As an example of a simple setup in which rationality restrictions and economic behavior may be analyzed, we take consumer demand. We establish that the relationship between nonparametric IV models and the recently emphasized issue of unobserved heterogeneity in structural models is tenuous at best for the case of consumer demand. As such, we believe that in such a problem a nonseparable control function approach is perhaps more suitable to deal with the issues of endogeneity and preference heterogeneity (see, e.g., Lewbel, 2001; Hoderlein, 2008). Still, the fact that approaches like the one put forward in this paper dispense with the requirement in the control function models that one has to specify the relationship that defines the control functions (i.e., the relationship

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

between endogeneous regressors and instruments) makes these models attractive. Therefore, approaches involving nonparametric IV that lead to ill-posed inverse problems may have some scope in demand analysis if one is not too concerned about unobserved heterogeneity.

The obvious following question is then: How do these models perform in practice? As in many applications that involve continuous endogenous regressors and instruments, it is also the case in consumer demand that these are approximately normally distributed. Unfortunately, this is bad news for the literature on inverse problems, since a normal specification leads to a severely ill-posed inverse problem with very slow (in general only logarithmic) convergence rates. Moreover, an additional error is introduced when estimating the operator. To achieve the best possible performance and to construct a simple and transparent estimator, in this paper we show how knowledge of joint (conditional) normality may be incorporated when estimating these models semiparametrically, and we devise and analyze such an estimator. An interesting feature of our semiparametric approach is that it suggests a test for identification.

Both in our simulation, as well as in the application, we do find our estimator to perform reasonably well, at least when one takes into account that it is designed to provide a good fit in the center of the data (in the tails, the behavior is less satisfactory). However, the estimator turns out to be quite sensitive toward specification issues, like the choice of the truncation parameter, and to misspecification (e.g., the breakdown when wrongly assuming joint instead of conditional normality). In particular, higher order Hermite polynomials in particular exhibit a high degree of variability. Choosing other basis functions may improve the performance of the estimator in finite sample situations, but at the cost of a more sophisticated estimator, and the precise performance still has to be investigated in our setup with continuous exogenuous covariates.

In summary, this paper has shown that estimation methods work reasonably well even in severely ill-posed problems, and as such may be a useful tool in those applications where the structural model is indeed given by a nonparametric model with additive error.

NOTES

1. Of course, joint (conditional) normality can be tested, and we will propose and implement a test below. A byproduct is that once conditional normality is accepted, it is actually possible to derive a test for identification, as demonstrated below.

2. They consider an L-vector Y, but we restrict ourselves to the scalar case. For the point we raise, this is immaterial.

3. We could consider the median or other quantities. Unsurprisingly, this would only make matters worse.

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

REFERENCES

Aït-Sahalia, Y., P.J. Bickel, & T.M. Stoker (2001) Goodness-of-fit tests for kernel regression with an application to option implied volatilities. *Journal of Econometrics* 105, 363–412.

Blondin, D. (2007) Rates of strong uniform convergence for local least squares kernel regression estimators. *Statistics & Probability Letters* 77, 1526–1534.

Blundell, R., X. Chen, & D. Kristensen (2007) Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* 75, 1613–1669.

Brown, B. & I. Walker (1989) The random utility hypothesis and inference in demand systems. *Econometrica* 57, 815–829.

Carrasco, M., J.-P. Florens, & E. Renault (2005) Linear inverse problems in structural econometrics: Estimation based on spectral decomposition and regularization. In J.J. Heckman & E.E. Leamer (eds.), *Handbook of Econometrics,* vol. 6, pp. 5633–5751. Elsevier.

Cavalier, L. & A. Tsybakov (2002) Sharp adaptation for inverse problems with random noise, *Probability Theory and Related Fields* 123, 323–354.

Darolles, S., J.-P. Florens, & E. Renault (2006) Nonparametric Instrumental Regression. Working Paper, Toulouse.

Delgado, M.A. & W. Stute (2008) Distribution-free specification tests of conditional models. *Journal of Econometrics* 143, 37–55.

Donoho, D.L. (1995) Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied and Computational. Harmonic Analyses* 2, 101–126.

Erdelyi, A. (1939) Über eine erzeugende funktion von produkten Hermitischer polynome (German). *Mathematische zeitsehsitt* 44, 201–210.

Fan, J. (1992) Design-adaptive nonparametric regression. *Journal of the Americian Statistical Association* 87, 998–1004.

Hall, P. & J.L. Horowitz (2005) Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics* 33, 2904–2929.

Härdle, W. & E. Mammen (1993) Comparing nonparametric versus parametric regression fits, *Annals of Statistics* 21, 1926–1947.

Hildenbrand, W. (1993) *Market Demand: Theory and Empirical Evidence*. Princeton University Press.

Hoderlein, S. (2008) Nonparametric Demand Systems, Instrumental Variables and a Heterogeneous Population. Working paper, Mannheim.

Hoderlein, S., Y. Klemelä, & E. Mammen (2007) Reconsidering the Random Coefficients Model. Working paper, Mannheim.

Hoderlein, S. & E. Mammen (2007) Identification of marginal effects in nonseparable models without monotonicity. *Econometrica* 75, 1513–1518.

Imbens, G. & W. Newey (2007) Identification and Estimation of Triangular Simultaneous Equations Models without Additivity. Working paper, MIT.

Johnstone, I.M. & B.W. Silverman (1990) Speed of estimation in positron emission tomography and related inverse problems. *Annals of Statistics* 18, 251–280.

Kress, R. (1989) *Linear Integral Equations*. Springer.

Lewbel, A. (2001) Demand systems with and without errors. *American Economic Review* 91, 611–618.

Mair, B.A. & F. Ruymgaart (1996) Statistical inverse estimation in Hilbert scales. *SIAM Journal Applied Mathematics* 56, 1424–1444.

Neumann, M.H. (1994) Pointwise confidence intervals in nonparametric regression with heteroscedastic error structure. *Statistics* 29, 1–36.

Newey, W.K. & J.L. Powell (2003) Instrumental variable estimation of nonparametric models. *Econometrica* 71, 1565–1578.

Nychka, D.W. & D. Cox (1989) Convergence rates for generalized solutions of integral equations from discrete noisy data. *Annals of Statistics* 17, 556–572.

Pendakur, K. (1999) Semiparametric estimates and tests of base-independent equivalence scales. *Journal of Econometrics* 88, 1–40

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

O'Sullivan, F. (1986) A statistical perspective on ill-posed inverse problems. *Statistical Science* 4, 169–184.

Severini, T.A. & G. Tripathi (2006) Some identification issues in nonparametric linear models with endogenous regressors. *Econometric Theory* 22, 258–278.

Szegö, G. (1959) *Orthogonal polynomials*, Rev. ed. American Mathematical Society.

APPENDIX: Proofs

For simplicity, we give the proof of Theorem 2 only for the conditionally normalized case (2.6). Estimating the error when standardizing is detailed in the proof of Theorem 3, and the case of Theorem 2 is quite similar. First we prove two lemmas.

LEMMA 1. *Under Assumptions 2 and 3, we have that*

$$\Big(\sup_{z\in I}\sum_{k=0}^{M}\Big|\frac{1}{\hat\rho^k(z)}-\frac{1}{\rho^k(z)}\Big|^2\Big)=O_P\Big(M^2(2/\rho_{\min})^{2(M+1)}n^{-2\epsilon_0}\Big). \tag{A.1}$$

where $\rho_{\min}$ *is as in Assumption 2. Further,*

$$\Big(\sup_{z\in I}\sum_{k=0}^{M}\Big|\frac{1}{\hat\rho^k(z)}-\frac{1}{\rho^k(z)}\Big|\Big)=O_P\Big(M(2/\rho_{\min})^{M+1}n^{-\epsilon_0}\Big). \tag{A.2}$$

Proof. From Assumptions 2 and 3, given $\delta>0$ there is an n_0 such that for all $n\ge n_0$,

$$\mathbb{P}\big(|\hat\rho(z)-\rho(z)|>|\rho(z)/2|\forall z\in I\big)<\delta/2.$$

Hence

$$\mathbb{P}\big(|\hat\rho(z)|>\rho_{\min}/2\forall z\in I\big)<\delta/2. \tag{A.3}$$

Now, we have that $\hat\rho^{-k}(z)-\rho^{-k}(z)=-k\xi_{k,z}^{-k+1}(\hat\rho(z)-\rho(z))$, where $\xi_{k,z}$ is between $\hat\rho(z)$ and $\rho(z)$. Since $|\rho(z)|>\rho_{\min}$, from (A.3) we get for $n\ge n_0$ with probability $>1-\delta/2$,

$$\Big|\frac{1}{\hat\rho^k(z)}-\frac{1}{\rho^k(z)}\Big|^2\le k^2(\rho_{\min}/2)^{-2(k+1)}|\hat\rho(z)-\rho(z)|^2.$$

From Assumption 3, given $\eta>0$ we can choose n_0 so large that for $n\ge n_0$, we also have

$$\mathbb{P}\big(|\hat\rho(z)-\rho(z)|^2>\eta/n^{2\epsilon_0}\forall z\in I\big)<\delta/2.$$

Using $\sum_{k=1}^{M}k^2x^k\le C\,M^2x^{M+1}$ for some $C>0$ we conclude that for $n\ge n_0$,

$$\mathbb{P}\Big(\sum_{k=0}^{M}\Big|\frac{1}{\hat\rho^k(z)}-\frac{1}{\rho^k(z)}\Big|^2>C\eta\,M^2(2/\rho_{\min})^{2(M+1)}n^{-2\epsilon_0}\forall z\in I\Big)<\delta,$$

which proves (A.1). The proof of (A.2) proceeds along completely similar lines. ■

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

LEMMA 2. *Under the assumptions of Theorem 2, there is a $C > 0$ such that for any $k \geq 0$ and any $z \in I$,*

$$\mathbb{E}\big(\hat{\beta}_k(z) - \beta_k(z)\big)^2 \leq Cn^{-4/5}. \tag{A.4}$$

Proof. Using Assumptions 1 and 2 it is easy to show that the regression functions $\beta_k(z) = \rho^k(z)\alpha_k(z)$ have uniformly bounded second derivatives. When estimating $\beta_k(z)$, note that the distribution of the covariate Z is the same for all k. The conditional variance does depend on k, but it will be enough to show that it is uniformly bounded (in z and k). To this end, we estimate

$$\begin{aligned}\mathrm{Var}\big(\mathrm{Y}^2\mathrm{H}_\mathrm{k}^2(\mathrm{W})|\mathrm{Z}=\mathrm{z}\big) &\leq \mathbb{E}\big(Y^2 H_k^2(W)|Z=z\big)\\ &\leq 2\Big(\mathbb{E}\big(m^2(X,Z)\,H_k^2(W)|Z=z\big)\\ &\qquad + \mathbb{E}\big(U^2 H_k^2(W)|Z=z\big)\Big).\end{aligned}$$

Since m is uniformly bounded by Assumption 2, and we are under the conditionally normalized case,

$$\mathbb{E}\big(m^2(X,Z)\,H_k^2(W)|Z=z\big) \leq \|m^2\|_\infty \mathbb{E}\big(H_k^2(W)|Z=z\big) = \|m^2\|_\infty,$$

and

$$\mathbb{E}\big(U^2 H_k^2(W)|Z=z\big) \leq \|\sigma^2\|_\infty \mathbb{E}\big(H_k^2(W)|Z=z\big) = \|\sigma^2\|_\infty.$$

Therefore, using results of Fan (1992, Thm. 1 and Rmk. 1), under Assumptions 4 and 5 we have for the pointwise MSE that

$$\mathbb{E}\big(\hat{\beta}_k(z) - \beta_k(z)\big)^2 \leq Cn^{-4/5}$$

where $C > 0$ can be chosen independently of k and z. ∎

Proof of Theorem 2. First we prove (2.11). We have

$$\begin{aligned}\|\hat{m}_{sco}(\cdot,z) - m(\cdot,z)\|_{f_{X|Z=z}}^2 &= \Big\| \sum_{k=0}^{M} \frac{\hat{\beta}_k(z)}{\hat{\rho}^k(z)} H_k(\cdot) - \sum_{k\geq 0} \alpha_k(z) H_k(\cdot) \Big\|_{f_{X|Z=z}}\\ &= \sum_{k=0}^{M} \Big| \frac{\hat{\beta}_k(z)}{\hat{\rho}^k(z)} - \frac{\beta_k(z)}{\rho^k(z)} \Big|^2 + \sum_{k\geq M+1} |\alpha_k(z)|^2. \end{aligned}\tag{A.5}$$

The bias term $\sum_{k\geq M+1} |\alpha_k(z)|^2$ can be easily estimated using Assumption 1 as $O(M^{-2\gamma+1})$. Now let us turn to the variance term in (A.5). We have

$$\sum_{k=0}^{M} \Big| \frac{\hat{\beta}_k(z)}{\hat{\rho}^k(z)} - \frac{\beta_k(z)}{\rho^k(z)} \Big|^2 \leq 2 \sum_{k=0}^{M} |\hat{\beta}_k(z)|^2 \Big| \frac{1}{\hat{\rho}^k(z)} - \frac{1}{\rho^k(z)} \Big|^2 + 2 \sum_{k=0}^{M} \Big| \frac{\hat{\beta}_k(z) - \beta_k(z)}{\rho^k(z)} \Big|^2. \tag{A.6}$$

In order to bound the second term on the right in (A.6), we estimate by using (A.4) that for any $z \in I$,

$$\begin{aligned}\mathbb{E}\Big(\sum_{k=0}^{M} \Big| \frac{\hat{\beta}_k(z) - \beta_k(z)}{\rho^k(z)} \Big|^2 \Big) &\leq \rho_{\min}^{-2M} \sum_{k=0}^{M} \mathbb{E}\big|\hat{\beta}_k(z) - \beta_k(z)\big|^2\\ &\leq C\,M\,n^{-4/5}\,\rho_{\min}^{-2M},\end{aligned}$$

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

where ρ is as in Assumption 2 and $C > 0$ does not depend on z. Thus, we also get convergence in probability with constant not depending on z. In order to bound the first term on the right in (A.6), we estimate

$$\sum_{k=0}^{M} |\hat{\beta}_k(z)|^2 \Big| \frac{1}{\hat{\rho}^k(z)} - \frac{1}{\rho^k(z)} \Big|^2 \le \max_{k=0}^{M} |\hat{\beta}_k(z)|^2 \sum_{k=0}^{M} \Big| \frac{1}{\hat{\rho}^k(z)} - \frac{1}{\rho^k(z)} \Big|^2.$$

We have

$$\begin{aligned} \Big| \max_{k=0}^{M} \hat{\beta}_k(z)^2 - \max_{k=0}^{M} \beta_k(z)^2 \Big| &\le \Big| \max_{k=0}^{M} \big(\hat{\beta}_k(z)^2 - \beta_k(z)^2 \big) \Big| \\ &\le \max_{k=0}^{M} \big| \hat{\beta}_k(z) - \beta_k(z) \big| \max_{k=0}^{M} \big| \hat{\beta}_k(z) + \beta_k(z) \big|. \end{aligned}$$

Further,

$$\sup_{z \in I} \mathbb{E} \Big(\max_{k=0}^{M} |\hat{\beta}_k(z) - \beta_k(z)|^2 \Big) \le \sup_{z \in I} \sum_{k=0}^{M} \mathbb{E} |\hat{\beta}_k(z) - \beta_k(z)|^2 \le C\, M\, n^{-4/5}.$$

Therefore, we get $\max_{k=0}^{M} |\hat{\beta}_k(z) - \beta_k(z)|^2 = O_P(M n^{-4/5})$, where the constant in O_P does not depend on z, and therefore also

$$\max_{k=0}^{M} |\hat{\beta}_k(z) - \beta_k(z)| = O_P(M^{1/2} n^{-2/5}),$$

where the constant in O_P does not depend on z. Moreover, since the $\beta_k(z)$ are bounded uniformly in k and z, we furthermore get that

$$\max_{k=0}^{M} |\hat{\beta}_k(z) + \beta_k(z)| = O_P(1).$$

Combining these estimates and (A.1), we get that

$$\sum_{k=0}^{M} |\hat{\beta}_k(z)|^2 \Big| \frac{1}{\hat{\rho}^k(z)} - \frac{1}{\rho^k(z)} \Big|^2 = O_P(1) O_P\Big(M^2 (2/\rho_{\min})^{2(M+1)} n^{-2\epsilon_0} \Big),$$

where the constants do not depend on z. Therefore, we get that

$$\begin{aligned} \|\hat{m}_{sco}(\cdot, z) - m(\cdot, z)\|^2_{f_{X|Z=z}} &= O_P\big(n^{-\min(4/5, 2\epsilon_0)} M^2 (2/\rho_{\min})^{2(M+1)} \big) \\ &\quad + O(M^{-2\gamma+1}), \end{aligned} \tag{A.7}$$

where the constants in O_P and O do not depend on z. Letting $M = c \log n$ for sufficiently small c, we get (2.11).

The proof of (2.12) proceeds along similar lines. We start with

$$\begin{aligned} \big| \hat{m}_{sco}(x, z) - m(x, z) \big| &\le \Big| \sum_{k=0}^{M} \frac{\hat{\beta}_k(z)}{\hat{\rho}^k(z)} H_k(x) - \sum_{k \ge 0} \frac{\beta_k(z)}{\rho^k(z)} H_k(x) \Big| \\ &\le \sum_{k=0}^{M} \Big| \frac{\hat{\beta}_k(z)}{\hat{\rho}^k(z)} - \frac{\beta_k(z)}{\rho^k(z)} \Big| |H_k(x)| + \sum_{k \ge M+1} \big| \alpha_k(z) \big| |H_k(x)|. \end{aligned} \tag{A.8}$$

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

On any compact subset, the $H_k(x)$ are uniformly bounded. In fact, one even has that

$$|H_k(x)| \leq C(J)(k+1)^{-1/4}, \qquad x \in J, \qquad k \geq 0,$$

cf. Szegö (1959, p. 242), and note that this also holds for the Hermite polynomials (not only the Hermite functions), since $e^{x^2/2}$ is also bounded on a compact set J. Therefore, using Assumption 1, the second term in (A.8) can be bounded by

$$\sum_{k \geq M+1} \left|\alpha_k(z)\right| |H_k(x)| \leq C(J) \sum_{k \geq M+1} k^{-(\gamma+1/4)} \leq C M^{-\gamma+3/4}.$$

Furthermore, arguing as above by using (A.2) one can show that

$$\begin{aligned}\sum_{k=0}^{M} \left| \frac{\hat{\beta}_k(z)}{\hat{\rho}^k(z)} - \frac{\beta_k(z)}{\rho^k(z)} \right| |H_k(x)| &\leq C(J) \sum_{k=0}^{M} \left| \frac{\hat{\beta}_k(z)}{\hat{\rho}^k(z)} - \frac{\beta_k(z)}{\rho^k(z)} \right| \\ &= O_P\left(n^{-\min(2/5,\epsilon_0)} M \left(2/\rho_{\min}\right)^{M+1}\right),\end{aligned}$$

where the constant in O_P does not depend on z. Therefore, for any $z \in I$,

$$\max_{x \in J} \left|\hat{m}_{sco}(x,z) - m(x,z)\right| = O_P\left(n^{-\min(2/5,\epsilon_0)} M \left(2/\rho_{\min}\right)^{M+1}\right) + O\left(M^{-\gamma+3/4}\right),$$

where the constants in O_p and O do not depend on z. Letting once more $M = c \log n$ for sufficiently small c gives (2.12). ■

Proof of Theorem 3. Let $\tilde{m}_{JN}(x,z)$ denote the estimator (2.15), where the true values for the parameters σ_1, σ_2, and ρ are used. We first bound the difference between this estimator and m, even in expectation. Assume $\rho > 0$. Compute

$$\begin{aligned}\mathbb{E}\|\tilde{m}_{JN} - m\|^2_{L_2(\mu_{XZ})} &= \sum_{j \geq 0} \sum_{k \geq M_2+1} |\alpha_{j,k}|^2 + \sum_{j \geq M_1} \sum_{k \leq M_2} |\alpha_{j,k}|^2 \\ &\quad + \frac{1}{n} \sum_{j=0}^{M_1} \sum_{k=0}^{M_2} \frac{\mathbb{E}_1^2 H_j^2(W_1 - \rho_{WZ} Z_1) H_k^2(Z_1) - \beta_{j,k}^2}{\rho^{2j}}.\end{aligned} \tag{A.9}$$

First consider the bias terms in (A.9). These can be estimated by using Assumption 6, as follows:

$$\sum_{j \geq 0} \sum_{k \geq M_2+1} |\alpha_{j,k}|^2 + \sum_{j \geq M_1} \sum_{k \leq M_2} |\alpha_{j,k}|^2 = O\left(M_2^{-2\delta+1} + M_1^{-2\gamma+1}\right).$$

Using Assumption 7, we have for the variance

$$\begin{aligned}&\mathbb{E} Y_1^2 H_j^2(W_1 - \rho_{WZ} Z_1) H_k^2(Z_1) - \beta_{j,k}^2 \\ &\quad \leq 2\mathbb{E}\Big(\left(m^2(X_1, Z_1) + U_1^2\right) H_j^2(W_1 - \rho_{WZ} Z_1) H_k^2(Z_1)\Big) \\ &\quad \leq 2\left(\|m\|_\infty^2 + \sigma^2\right).\end{aligned}$$

Hence

$$\begin{aligned}\frac{1}{n} \sum_{j=0}^{M_1} \sum_{k=0}^{M_2} \frac{\mathbb{E} Y_1^2 H_j^2(W_1 - \rho_{WZ} Z_1) H_k^2(Z_1) - \beta_{j,k}^2}{\rho^{2j}} &\leq \frac{2}{n} \sum_{j=0}^{M_1} \sum_{k=0}^{M_2} \frac{\|m\|_\infty^2 + \|\sigma^2\|_\infty}{\rho^{2j}} \\ &\leq \frac{C}{n} (M_2+1) \rho^{-2(M_1+1)}.\end{aligned}$$

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

Hence, we get that

$$\mathbb{E}\|\tilde{m}_{JN} - m\|^2_{L_2(\mu_{XZ})} = O\left(n^{-1}(M_2+1)\rho^{-2(M_1+1)} + M_2^{-2\delta+1} + M_1^{-2\gamma+1}\right).$$

Letting $M_1 = c\log n$ for sufficiently small $c > 0$ and $M_2 = n^{-\epsilon}$ for small ϵ gives the rate in (2.16), since the term $M_1^{-2\gamma+1}$ dominates.

Now we bound the difference between $\hat{m}_{JN}$ and $\tilde{m}_{JN}$ using Assumption 8. Start with

$$\|\hat{m}_{JN} - \tilde{m}_{JN}\|^2_{L_2(\mu_{XZ})} \le 2\Big\|\sum_{j\le M_1}\sum_{k\le M_2}\frac{\hat{\beta}_{j,k}}{\hat{\rho}^j}(\hat{\psi}_{j,k} - \psi_{j,k})\Big\|^2_{L_2(\mu_{XZ})} \tag{A.10}$$
$$+2\Big\|\sum_{j\le M_1}\sum_{k\le M_2}\Big(\frac{\hat{\beta}_{j,k}}{\hat{\rho}^j} - \frac{\beta_{j,k}}{\rho^j}\Big)\psi_{j,k}\Big\|^2_{L_2(\mu_{XZ})}.$$

We start by estimating the second term. Compute

$$\Big\|\sum_{j\le M_1}\sum_{k\le M_2}\Big(\frac{\hat{\beta}_{j,k}}{\hat{\rho}^j} - \frac{\beta_{j,k}}{\rho^j}\Big)\psi_{j,k}\Big\|^2_{L_2(\mu_{XZ})} = \sum_{j\le M_1}\sum_{k\le M_2}\Big|\frac{\hat{\beta}_{j,k}}{\hat{\rho}^j} - \frac{\beta_{j,k}}{\rho^j}\Big|^2$$
$$\le 2\sum_{j\le M_1}\sum_{k\le M_2}\Big|\frac{\hat{\beta}_{j,k} - \beta_{j,k}}{\hat{\rho}^j}\Big|^2 + 2\sum_{j\le M_1}\sum_{k\le M_2}\beta_{j,k}^2\Big|\frac{1}{\hat{\rho}^j} - \frac{1}{\rho^j}\Big|^2.$$

The second term is estimated similarly as in Lemma 1 as follows:

$$\Big\|\sum_{j\le M_1}\sum_{k\le M_2}\Big(\frac{\hat{\beta}_{j,k}}{\hat{\rho}^j} - \frac{\beta_{j,k}}{\rho^j}\Big)\psi_{j,k}\Big\|^2_{L_2(\mu_{XZ})} = O_P\big(M_2^2(2/\rho)^{2(M_1+1)}n^{-1}\big)$$
$$= O_P\big(n^{-\delta}\big).$$

For the first term, we estimate using $H_j'(x) = j\,H_{j-1}(x)$ that

$$|\hat{\beta}_{j,k} - \beta_{j,k}| = \Big|\frac{1}{n}\sum_{i=1}^n Y_i H_k(Z_i)\Big(H_j\big(\frac{W_i - \hat{\rho}_{WZ}Z_i}{\hat{\sigma}_2}\big) - H_j\big(\frac{W_i - \rho_{WZ}Z_i}{\sigma_2}\big)\Big)\Big|$$
$$= \Big|\frac{1}{n}\sum_{i=1}^n Y_i H_k(Z_i)\,j\,H_{j-1}(\xi_i)Z_i\Big(\frac{\rho_{WZ}}{\sigma_2} - \frac{\hat{\rho}_{WZ}}{\hat{\sigma}_2}\Big)\Big|$$
$$= O_P\big(jn^{-1/2}\big),$$

where the intermediate values ξ_i can be chosen identically distributed. Hence,

$$\sum_{j\le M_1}\sum_{k\le M_2}\Big|\frac{\hat{\beta}_{j,k} - \beta_{j,k}}{\hat{\rho}^j}\Big|^2 = O_P\big(M_1^3(2/\rho)^{2(M_2+1)}n^{-1}\big) = O_P(n^{-\delta}).$$

Now we consider the first term in (A.10). We have

$$\Big\|\sum_{j\le M_1}\sum_{k\le M_2}\frac{\hat{\beta}_{j,k}}{\hat{\rho}^j}(\hat{\psi}_{j,k} - \psi_{j,k})\Big\|^2_{L_2(\mu_{XZ})}$$
$$\le \Big(\sum_{j\le M_1}\sum_{k\le M_2}\Big|\frac{\hat{\beta}_{j,k}}{\hat{\rho}^j}\Big|\big\|\hat{\psi}_{j,k} - \psi_{j,k}\big\|_{L_2(\mu_{XZ})}\Big)^2.$$

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press

Estimate

$$\begin{aligned}\|\hat{\psi}_{j,k} - \psi_{j,k}\|^2_{L_2(\mu_{XZ})} &= \int\int \left| H_j\big((x - \hat{\rho}_{XZ} z)/\hat{\sigma}_1\big) - H_j\big((x - \rho_{XZ} z)/\sigma_1\big)\right|^2 \\ &\quad \times |H_k(z)|^2 \, d\mu_{XZ} \\ &= \int\int \left| j H_{j-1}(\xi_{x,z}) \frac{\rho_{XZ}}{\sigma_1} - \frac{\hat{\rho}_{XZ}}{\hat{\sigma}_1} z^2 |H_k(z)|^2 \, d\mu_{XZ} \right. \\ &= O_P(j^2 n^{-1}).\end{aligned}$$

Therefore

$$\Big\| \sum_{j \le M_1} \sum_{k \le M_2} \frac{\hat{\beta}_{j,k}}{\hat{\rho}^j} (\hat{\psi}_{j,k} - \psi_{j,k}) \Big\|^2_{L_2(\mu_{XZ})} = O_P\big(M_1^4 (2/\rho)^{2(M_1+1)} M_2^2 n^{-1}\big).$$

Collecting terms, this concludes the proof of Theorem 3. ■

https://doi.org/10.1017/S0266466610000423 Published online by Cambridge University Press