OPTIMAL SUP-NORM RATES, ADAPTIVITY AND INFERENCE IN NONPARAMETRIC INSTRUMENTAL VARIABLES ESTIMATION

By

Xiaohong Chen and Timothy Christensen

November 2013
Revised April 2015

COWLES FOUNDATION DISCUSSION PAPER NO. 1923R

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Optimal Sup-norm Rates, Adaptivity and Inference in Nonparametric Instrumental Variables Estimation*

Xiaohong Chen† and Timothy M. Christensen‡

First version August 2013; Revised March 31, 2015

Abstract

This paper makes several contributions to the literature on the important yet difficult problem of estimating functions nonparametrically using instrumental variables. First, we derive the minimax optimal sup-norm convergence rates for nonparametric instrumental variables (NPIV) estimation of the structural function h_0 and its derivatives. Second, we show that a computationally simple sieve NPIV estimator can attain the optimal sup-norm rates for h_0 and its derivatives when h_0 is approximated via a spline or wavelet sieve. Our optimal sup-norm rates surprisingly coincide with the optimal L^2-norm rates for severely ill-posed problems, and are only up to a $[\log(n)]^\epsilon$ (with $\epsilon < 1/2$) factor slower than the optimal L^2-norm rates for mildly ill-posed problems. Third, we introduce a novel data-driven procedure for choosing the sieve dimension optimally. Our data-driven procedure is sup-norm rate-adaptive: the resulting estimator of h_0 and its derivatives converge at their optimal sup-norm rates even though the smoothness of h_0 and the degree of ill-posedness of the NPIV model are unknown. Finally, we present two non-trivial applications of the sup-norm rates to inference on nonlinear functionals of h_0 under low-level conditions. The first is to derive the asymptotic normality of sieve t-statistics for exact consumer surplus and deadweight loss functionals in nonparametric demand estimation when prices, and possibly incomes, are endogenous. The second is to establish the validity of a sieve score bootstrap for constructing asymptotically exact uniform confidence bands for collections of nonlinear functionals of h_0. Both applications provide new and useful tools for empirical research on nonparametric models with endogeneity.

Keywords: Ill-posed inverse problems; Series 2SLS; Optimal sup-norm convergence rates; Adaptive estimation; Random matrices; Bootstrap uniform confidence bands; Nonlinear welfare functionals; Nonparametric demand analysis with endogeneity.

---

*This paper is a major extension of Sections 2 and 3 of our Cowles Foundation Discussion Paper CFDP1923, Cemmap Working Paper CWP56/13 and arXiv preprint arXiv:1311.0412 [math.ST] (Chen and Christensen, 2013). We thank E. Gautier, E. Guerre, J. Horowitz, S. Lee, W. Newey, J. Powell, A. Tsybakov and participants of SETA2013, AMES2013, SETA2014, the 2014 International Symposium in honor of Jerry Hausman, the 2014 Cowles Summer Conference, the 2014 SJTU-SMU Econometrics Conference, the 2014 Cemmap Celebration Conference, and seminars at Caltech, Johns Hopkins, NYU, Toulouse, and USC for useful comments. Support from the Cowles Foundation is gratefully acknowledged. Any errors are the responsibility of the authors.

†Cowles Foundation for Research in Economics, Yale University, Box 208281, New Haven, CT 06520, USA. E-mail address: xiaohong.chen@yale.edu

‡Department of Economics, New York University, 19 W. 4th St, 6th floor, New York, NY 10012, USA. E-mail address: timothy.christensen@nyu.edu

1 Introduction

This paper investigates how well one may estimate an unknown structural function h_0 of endogenous regressors in sup-norm loss, where h_0 is identified by a nonparametric instrumental variables (NPIV) model: $E[Y_i - h_0(X_i)|W_i] = 0$, where X_i is a vector of endogenous regressors and W_i is a vector of instrumental variables. We show that a computationally simple sieve NPIV estimator, which is also called a sieve minimum distance or series 2SLS estimator (Newey and Powell, 2003; Ai and Chen, 2003; Blundell, Chen, and Kristensen, 2007), can attain the best possible sup-norm convergence rates when spline or wavelet sieves are used to approximate the unknown h_0. We introduce a novel data-driven procedure to choose the key regularization parameter, namely the sieve dimension (for approximating h_0), of the sieve NPIV estimator. The procedure is shown to be sup-norm rate adaptive to the unknown smoothness of h_0 and the unknown degree of ill-posedness of the NPIV model. In fact, we show that the same data-driven choice of the sieve dimension simultaneously leads to the optimal sup-norm rates for estimating h_0 and its derivatives.

Sup-norm (uniform) convergence rates for nonparametric estimators of h_0 and its derivatives provide sharper measures on how well the unknown function h_0 and its derivatives could be estimated given a sample of size n. Equally or perhaps more importantly, they are very useful to control the nonlinearity bias when conducting inference on nonlinear functionals of h_0, such as the exact consumer surplus and deadweight loss welfare functionals in nonparametric demand estimation (Hausman and Newey, 1995; Vanhems, 2010; Blundell, Horowitz, and Parey, 2012).

NPIV estimation has been the subject of much recent research, both because of its importance to applied economics and its prominent role in the literature on ill-posed inverse problems with unknown operators. In addition to the sieve NPIV estimator (Newey and Powell, 2003; Ai and Chen, 2003; Blundell et al., 2007), other estimators have also been considered in the literature; see Hall and Horowitz (2005); Carrasco, Florens, and Renault (2007); Darolles, Fan, Florens, and Renault (2011); Horowitz (2011); Liao and Jiang (2011); Gagliardini and Scaillet (2012); Chen and Pouzo (2012); Florens and Simoni (2012); Kato (2013) and references therein. To the best of our knowledge, all the published works on convergence rates for various NPIV estimators have only studied L^2-norm (or closely related Hilbert-norm) convergence rates. In particular, Hall and Horowitz (2005) are the first to establish the minimax lower bound in L^2-norm loss for estimating h_0 for mildly ill-posed NPIV models, and show that their estimators attain the lower bound. Chen and Reiss (2011) derive the minimax lower bound in L^2-norm loss for estimating h_0 for NPIV models that could be mildly or severely ill-posed, and show that the sieve NPIV estimator achieves the lower bound.[1] Recently, for Horowitz (2011)'s modified orthogonal series NPIV estimator, Horowitz (2014) proposed a data-driven procedure for choosing the orthogonal series dimension that is near L^2-norm rate-adaptive in that his procedure attains the optimal L^2-norm rate up to a $[\log(n)]^{1/2}$

[1] Subsequently, some other NPIV estimators have also been shown to attain the optimal L^2-norm rates.

factor for both mildly or severely ill-posed models. As yet there are no published results on sup-norm convergence rates for *any* NPIV estimator, nor are there results on what are the minimax lower bounds in sup-norm loss for any class of NPIV models. Further, there is no prior work on any data-driven procedure that is sup-norm rate adaptive to the unknown smoothness of the function h_0 and the unknown degree of ill-posedness of the NPIV model.

In this paper we study the sup-norm convergence properties of the sieve NPIV estimator of the unknown h_0 and its derivatives. We focus on this estimator because, in addition to its known L^2-norm rate optimality for estimating h_0, it has been used extensively in empirical work and can be implemented as a simple two stage least squares (2SLS) estimator even when X_i contains both endogenous and exogenous regressors. We first establish a general upper bound on the sup-norm convergence rate of any sieve NPIV estimator. When h_0 belongs to a Hölder ball of functions with smoothness $p > 0$, we obtain the sup-norm convergence rates of the spline and wavelet sieve NPIV estimators for estimating h_0 and its derivatives jointly. We then derive the minimax lower bounds in sup-norm loss for h_0 and its derivatives uniformly over a Hölder ball of functions. The lower bounds are shown to equal our sup-norm convergence rates for the spline and wavelet sieve NPIV estimators of h_0 and its derivatives. Surprisingly, these optimal sup-norm convergence rates for estimating h_0 and its derivatives coincide with the optimal L^2-norm rates for severely ill-posed problems, and are a factor of $[\log(n)]^\epsilon$ (with $0 < \epsilon < p/(2p+1)$) slower than the optimal L^2-norm rates for mildly ill-posed problems.[2]

In practice, to attain the optimal sup-norm convergence rates of the sieve NPIV estimator one must choose the sieve dimension to balance the sup-norm bias term and the sup-norm sampling error term (loosely called the "standard deviation" term). The sup-norm bias term depends on the smoothness of the unknown function h_0 and the sup-norm standard deviation term depends on the degree of ill-posedness of the unknown NPIV operator. Therefore, it is important to have a method for choosing the optimal sieve dimension without knowing these unknowns. We introduce a new data-driven procedure for choosing the sieve dimension for approximating h_0. We show that our data-driven choice of sieve dimension is optimal in that the resulting sieve NPIV estimators of h_0 and its derivatives attain their optimal sup-norm rates. Interestingly, our data-driven procedure automatically leads to optimal L^2-norm rate adaptivity for severely ill-posed models, and optimal L^2-norm rate adaptivity up to a factor of $[\log(n)]^\epsilon$ (with $0 < \epsilon < p/(2p+1)$) for mildly ill-posed models. Our data-driven procedure is different from the model selection procedure proposed by Horowitz (2014) for his modified orthogonal series NPIV estimator, which might explain why our procedure leads to a L^2-norm rate that is faster than his procedure. A Monte Carlo study indicates that our sup-norm rate-adaptive procedure performs well in finite samples.

We illustrate the usefulness of the sup-norm convergence rate results with two non-trivial applica-

[2]See Subsection 2.4 for the expression of ϵ and the complementary results on L^2-norm rate optimality of sieve NPIV estimators for estimating the derivatives of h_0.

tions, each of which makes new contribution to inference on *nonlinear* welfare functionals in NPIV estimation. Inference on nonlinear functionals of h_0 in a NPIV model is very difficult because of the combined effects of nonlinearity bias and slow rates of convergence of any NPIV estimators in L^q norms for $1 \leq q \leq \infty$. In the first application, we extend the important work by Hausman and Newey (1995) on nonparametric estimation of exact consumer surplus and deadweight loss functionals to allow for prices, and possibly incomes, to be endogenous. Specifically, we use our sup-norm convergence rates for estimating the demand function h_0 and its derivatives to linearize plug-in estimators of exact consumer surplus and deadweight loss functionals. This linearization immediately leads to asymptotic normality of sieve t statistics for exact consumer surplus and deadweight loss using our pointwise limit theory in Appendix A.[3] Our second important application is to sieve score bootstrap uniform confidence bands for collections of nonlinear functionals of h_0; see Appendix B for details.

The rest of the paper is organized as follows. Section 2 establishes the optimal rates of convergence for estimating h_0 and its derivatives in a NPIV model. Section 3 introduces a data-driven sup-norm rate-adaptive procedure for sieve NPIV estimators. Section 4 provides an application to inference on exact consumer surplus and deadweight loss functionals in nonparametric demand estimation with endogeneity. Appendices A and B present low-level conditions for pointwise and bootstrap uniform limit theories for sieve t statistics of general nonlinear functionals of h_0 in a NPIV model respectively. The online appendix contains background materials on B-spline and wavelet sieve spaces (Appendix C), technical lemmas and all the proofs (Appendix D), and supplementary useful lemmas on random matrices (Appendix E).

Notation: Throughout we work on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. $\mathcal{A}^c$ denotes the complement of a measurable event $\mathcal{A} \in \mathcal{F}$. We abbreviate "with probability approaching one" to "wpa1", and say that a sequence of events $\{\mathcal{A}_n\} \subset \mathcal{F}$ holds wpa1 if $\mathbb{P}(\mathcal{A}_n^c) = o(1)$. For a random variable X we define the space $L^q(X)$ as the equivalence class of all measurable functions of X with finite qth moment if $1 \leq q < \infty$; when $q = \infty$ with some abuse of notation we take $L^\infty(X)$ to mean the set of all bounded measurable $f : \mathcal{X} \to \mathbb{R}$ endowed with the sup norm $\|f\|_\infty = \sup_x |f(x)|$. We let $\langle \cdot, \cdot \rangle_X$ denote the inner product on $L^2(X)$. For matrix and vector norms, $\|\cdot\|_{\ell^q}$ denotes the vector ℓ^q norm when applied to vectors and the operator norm induced by the vector ℓ^q norm when applied to matrices. If a and b are scalars we let $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$.

[3] By exploiting the sup-norm rates and the close form expression of the sieve NPIV estimator, Appendix A derives the pointwise asymptotic normality of sieve t statistics of nonlinear functionals of NPIV under lower-level conditions with faster growth of sieve dimension than those in Chen and Pouzo (2014) for functionals of general semi/nonparametric conditional moment restrictions.

2 Optimal sup-norm convergence rates

This section consists of several subsections. Subsection 2.1 outlines the NPIV model and the estimator. Subsections 2.2 and 2.3 first establish a general upper bound on the sup-norm convergence rates for any sieve NPIV estimator and then the minimax lower bound in sup-norm loss, allowing for both mildly and severely ill-posed problems. These results together lead to the optimal sup-norm convergence rates for spline and wavelet sieve NPIV estimators for estimating h_0 and its derivatives. Subsection 2.4 shows that the sieve NPIV estimator attains the optimal L^2-norm convergence rates for estimating h_0 and its derivatives under much weaker conditions. Finally Subsection 2.5 considers an extended NPIV model with endogenous and exogenous regressors.

2.1 The NPIV model, the estimator and the measure of ill-posedness

Throughout the paper the data $\{(Y_i, X_i, W_i)\}_{i=1}^n$ is assumed to be a random sample from the nonparametric instrumental variables (NPIV) model

$$\begin{aligned} Y_i &= h_0(X_i) + u_i \\ E[u_i|W_i] &= 0, \end{aligned} \tag{1}$$

where $Y_i \in \mathbb{R}$ is a scalar response variable, $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ is a d-dimensional endogenous regressor, and $W_i \in \mathcal{W} \subseteq \mathbb{R}^{d_w}$ is a d_w vector of (conditional) instrumental variables.

The sieve NPIV (or series 2SLS) estimator $\widehat{h}$ of h_0 may be written in matrix form as

$$\widehat{h}(x) = \psi^J(x)'\widehat{c} \quad \text{with} \quad \widehat{c} = [\Psi' B (B'B)^- B'\Psi]^- \Psi' B (B'B)^- B'Y$$

where $Y = (Y_1, \ldots, Y_n)'$ and

$$\begin{aligned} \psi^J(x) &= (\psi_{J1}(x), \ldots, \psi_{JJ}(x))' & \Psi &= (\psi^J(X_1), \ldots, \psi^J(X_n))' \\ b^K(w) &= (b_{K1}(w), \ldots, b_{KK}(w))' & B &= (b^K(W_1), \ldots, b^K(W_n))' \end{aligned}$$

(see (Blundell et al., 2007; Newey, 2013)).

The crucial *regularization* parameter to be chosen is the dimension J of the sieve space used to approximate the structural function h_0. The *smoothing* parameter K is the dimension of the instrument space, and is assumed to grow at the order of $K = O(J)$. From the analogy with 2SLS, it is clear that we need $K \geq J$.[4] When $K = J$, $b^K = \psi^J$ and $d_w = d$, the sieve 2SLS estimator

[4]Previous Monte Carlo evidences (Blundell et al., 2007; Chen and Pouzo, 2014) have demonstrated that sieve NPIV estimators often perform better with $K > J$ (the "over identified" case) than with $K = J$ (the "just identified" case), and that the regularization parameter J is important for finite sample performance while the parameter K is not important as long as it is larger than J.

becomes Horowitz (2011)'s modified orthogonal series NPIV estimator.

The coefficient estimator $\widehat{c}$ is the 2SLS estimator of the parameter vector $c_{0,J} \in \mathbb{R}^J$ satisfying

$$E[b^K(W_i)(Y_i - \psi^J(X_i)'c_{0,J})] = 0\,.$$

Define

$$\begin{array}{rcccl} G_\psi & = & G_{\psi,J} & = & E[\psi^J(X_i)\psi^J(X_i)'] \\ G_b & = & G_{b,K} & = & E[b^K(W_i)b^K(W_i)'] \\ S & = & S_{JK} & = & E[b^K(W_i)\psi^J(X_i)']\,. \end{array}$$

We assume that S has full column rank J and that $G_{\psi,J}$ and $G_{b,K}$ are positive definite for each J and K, i.e., $e_J = \lambda_{\min}(G_{\psi,J}) > 0$ for each J and $e_{b,K} = \lambda_{\min}(G_{b,K}) > 0$ for each K. Then $c_{0,J}$ can be expressed as

$$c_{0,J} = [S'G_b^{-1}S]^{-1}S'G_b^{-1}E[b^K(W_i)Y_i] = [S'G_b^{-1}S]^{-1}S'G_b^{-1}E[b^K(W_i)h_0(X_i)]\,.$$

We refer to $\psi^J(\cdot)'c_{0,J}$ as the sieve 2SLS approximation to h_0.

To introduce a measure of ill-posedness, let $T : L^2(X) \to L^2(W)$ denote the conditional expectation operator given by

$$Th(w) = E[h(X_i)|W_i = w]\,.$$

The operator T is compact when X is endogenous under mild conditions on the conditional density of X given W (see, e.g., Newey and Powell (2003); Blundell et al. (2007); Darolles et al. (2011); Andrews (2011)). Let $\Psi_J = clsp\{\psi_{J1}, \ldots, \psi_{JJ}\} \subset L^2(X)$ and $B_K = clsp\{b_{K1}, \ldots, b_{KK}\} \subset L^2(W)$ denote the sieve spaces for the endogenous and instrumental variables, respectively.[5] Let $\Psi_{J,1} = \{h \in \Psi_J : \|h\|_{L^2(X)} = 1\}$. The *sieve L^2 measure of ill-posedness* is

$$\tau_J = \sup_{h \in \Psi_J : h \neq 0} \frac{\|h\|_{L^2(X)}}{\|Th\|_{L^2(W)}} = \frac{1}{\inf_{h \in \Psi_{J,1}} \|Th\|_{L^2(W)}}\,.$$

Following Blundell et al. (2007), we call a NPIV model:
(i) *mildly ill-posed* if $\tau_J = O(J^{\varsigma/d})$ for some $\varsigma > 0$; and
(ii) *severely ill-posed* if $\tau_J = O(\exp(\frac{1}{2}J^{\varsigma/d}))$ for some $\varsigma > 0$.

[5]The exception is when the vector of regressors contains both endogenous and exogenous variables. See Section 2.5 for a discussion.

2.2 Upper bounds on sup-norm rates

To derive the sup-norm (uniform) convergence rate we split $\|\widehat{h} - h_0\|_\infty$ into so-called bias and variance terms and derive sup-norm convergence rates for the two terms. Specifically, let

$$\widetilde{h}(x) = \psi^J(x)'\widetilde{c} \quad \text{with} \quad \widetilde{c} = [\Psi' B (B'B)^- B'\Psi]^- \Psi' B (B'B)^- B' H_0$$

where $H_0 = (h_0(X_1), \ldots, h_0(X_n))'$. We say that $\|\widetilde{h} - h_0\|_\infty$ is the "bias" term and $\|\widehat{h} - \widetilde{h}\|_\infty$ is the "standard deviation" (or sometimes loosely called "variance") term. Both are random quantities.

We first introduce some basic conditions on the supports of the data, identification, true residuals and the sieve spaces.

Assumption 1 *(i) X_i has compact rectangular support $\mathcal{X} \subset \mathbb{R}^d$ with nonempty interior and the density of X_i is uniformly bounded away from 0 and ∞ on $\mathcal{X}$; (ii) W_i has compact rectangular support $\mathcal{W} \subset \mathbb{R}^{d_w}$ and the density of W_i is uniformly bounded away from 0 and ∞ on $\mathcal{W}$; (iii) $T : L^2(X) \to L^2(W)$ is injective; and (iv) $h_0 \in \mathcal{H} \subset L^\infty(X)$ and $\Psi_J \subset \Psi_{J'}$ for $J' > J$ with $\cup_J \Psi_J$ dense in $(\mathcal{H}, \|\cdot\|_{L^2(X)})$.*

Assumption 2 *(i) $\sup_{w \in \mathcal{W}} E[u_i^2 | W_i = w] \leq \overline{\sigma}^2 < \infty$; and (ii) $E[|u_i|^{2+\delta}] < \infty$ for some $\delta > 0$.*

We say that the sieve basis for Ψ_J is Hölder continuous if there exist finite constants $\omega \geq 0, \omega' > 0$ such that $\|G_{\psi,J}^{-1/2}\{\psi^J(x) - \psi^J(x')\}\|_{\ell^2} \lesssim J^\omega \|x - x'\|_{\ell^2}^{\omega'}$ for all $x, x' \in \mathcal{X}$. Let

$$\zeta_\psi = \zeta_{\psi,J} = \sup_x \|G_\psi^{-1/2}\psi^J(x)\|_{\ell^2} \qquad \zeta_b = \zeta_{b,K} = \sup_w \|G_b^{-1/2} b^K(w)\|_{\ell^2}$$
$$\xi_\psi = \xi_{\psi,J} = \sup_x \|\psi^J(x)\|_{\ell^1}$$

for each J and K and define $\zeta = \zeta_J = \zeta_{b,K} \vee \zeta_{\psi,J}$. Note that $\zeta_{\psi,J}$ has some useful properties: $\|h\|_\infty \leq \zeta_{\psi,J}\|h\|_{L^2(X)}$ for all $h \in \Psi_J$, and $\sqrt{J} = (E[\|G_\psi^{-1/2}\psi^J(X)\|_{\ell^2}^2])^{1/2} \leq \zeta_{\psi,J} \leq \xi_{\psi,J}/\sqrt{e_J}$; clearly $\zeta_{b,K}$ has similar properties.

Assumption 3 *(i) the basis spanning Ψ_J is Hölder continuous; (ii) $\tau_J \zeta^2/\sqrt{n} = O(1)$; and (iii) $\zeta^{(2+\delta)/\delta}\sqrt{(\log n)/n} = o(1)$.*

Let $\Pi_K : L^2(W) \to B_K$ denote the $L^2(W)$ orthogonal projection onto B_K (the sieve instrumental variables space).

Assumption 4 *(i) $\sup_{h \in \Psi_{J,1}} \|(\Pi_K T - T)h\|_{L^2(W)} = o(\tau_J^{-1})$.*

Assumption 1 is standard. Parts (i) and (ii) just place some mild regularity conditions on the support of the data. Part (iii) is typically satisfied in models with endogeneity (e.g. Newey and Powell (2003); Carrasco et al. (2007); Blundell et al. (2007); Andrews (2011)). The parameter space $\mathcal{H}$ for h_0 in part (iv) is typically taken to be a Hölder or Sobolev class. Assumption 2(i)(ii) are also imposed for sup-norm convergence rates for series LS regression without endogeneity (e.g., Chen and Christensen (2014)). Assumption 3(i) is satisfied by most commonly used sieve bases. Assumption 3(ii)(iii) restrict the maximum rate at which J can grow with the sample size. Upper bounds for $\zeta_{\psi,J}$ and $\zeta_{b,K}$ are known for commonly used bases under standard regularity conditions. For instance, under Assumption 1(i)(ii), $\zeta_{b,K} = O(\sqrt{K})$ and $\zeta_{\psi,J} = O(\sqrt{J})$ for (tensor-product) polynomial spline, wavelet and cosine bases, and $\zeta_{b,K} = O(K)$ and $\zeta_{\psi,J} = O(J)$ for (tensor-product) orthogonal polynomial bases (see, e.g., Huang (1998) and online Appendix C). Assumption 4(i) is a very mild condition on the approximation properties of the basis used for the instrument space and is similar to the first part of Assumption 5(iv) of Horowitz (2014). It is trivially satisfied with $\|(\Pi_K T - T)h\|_{L^2(W)} = 0$ for all $h \in \Psi_J$ when the basis functions for B_K and Ψ_J form either a Riesz basis or eigenfunction basis for the conditional expectation operator.

2.2.1 Bound on sup-norm "standard derivation"

Lemma 2.1 *Let Assumptions 1(i)(iii), 2(i)(ii), 3(ii)(iii), and 4(i) hold. Then:*
(1) $\|\widehat{h} - \widetilde{h}\|_\infty = O_p\left(\tau_J \xi_{\psi,J} \sqrt{(\log n)/(ne_J)}\right)$.
(2) If Assumption 3(i) also holds, then: $\|\widehat{h} - \widetilde{h}\|_\infty = O_p\left(\tau_J \zeta_{\psi,J} \sqrt{(\log n)/n}\right)$.

Recall that $\sqrt{J} \leq \zeta_{\psi,J} \leq \xi_{\psi,J}/\sqrt{e_J}$. Result (2) of Lemma 2.1 provides a slightly tighter upper bound on the variance term than Result (1) does, while Result (1) allows for slightly more general basis to be used to approximate h_0. For splines and wavelets, we show in Appendix C that $\xi_{\psi,J}/\sqrt{e_J} \lesssim \sqrt{J}$, so Results (1) and (2) produce the same tight upper bound $\tau_J \sqrt{(J \log n)/n}$ on $\|\widehat{h} - \widetilde{h}\|_\infty$.

2.2.2 Bound on sup-norm "bias"

Before we provide a bound on the sup-norm "bias" term $\|\widetilde{h} - h_0\|_\infty$, we introduce various non-random projections of h_0 onto the sieve approximating space Ψ_J, which imply different sieve approximation errors for h_0 that have close relations among themselves.

Let $\Pi_J : L^2(X) \to \Psi_J$ denote the $L^2(X)$ orthogonal projection onto Ψ_J and then $\Pi_J h_0 = \arg\min_{h \in \Psi_J} \|h_0 - h)\|_{L^2(X)}$. Let $Q_J h_0 = \arg\min_{h \in \Psi_J} \|\Pi_K T(h_0 - h)\|_{L^2(W)}$ denote the sieve 2SLS projection of h_0 onto Ψ_J, which is $Q_J h_0 = \psi^J(\cdot)' c_{0,J}$. Let $\pi_J h_0 = \arg\min_{h \in \Psi_J} \|T(h_0 - h)\|_{L^2(W)}$ denote the IV projection of h_0 onto Ψ_J.

Assumption 4 (continued) *(ii)* $\tau_J \times \|T(h_0 - \Pi_J h_0)\|_{L^2(W)} \leq \text{const} \times \|h_0 - \Pi_J h_0\|_{L^2(X)}$*; and (iii)* $\|Q_J h_0 - \Pi_J h_0\|_\infty \leq O(1) \times \|h_0 - \Pi_J h_0\|_\infty$.

Assumption 4(ii) is the usual L^2 "stability condition" imposed in the NPIV literature (see Assumption 6 in Blundell et al. (2007) and Assumption 5.2(ii) in Chen and Pouzo (2012) and their sufficient conditions). Assumption 4(iii) is a new L^∞ "stability condition" to control for the sup-norm bias. Instead of Assumption 4(iii), we could impose the following Assumption 4(iii').

Assumption 4 *(iii')* $(\zeta_{\psi,J}\tau_J) \times \|(\Pi_K T - T)(Q_J h_0 - \pi_J h_0)\|_{L^2(W)} \leq \text{const} \times \|Q_J h_0 - \pi_J h_0\|_{L^2(X)}$.

Lemma 2.2 *Let Assumptions 1(iii) and 4(ii) hold. Then:*
(1) $\|h_0 - \pi_J h_0\|_{L^2(X)} \asymp \|h_0 - \Pi_J h_0\|_{L^2(X)}$*;*
(2) If Assumption 4(i) also holds, then: $\|Q_J h_0 - \pi_J h_0\|_{L^2(X)} \leq o(1) \times \|h_0 - \pi_J h_0\|_{L^2(X)}$.
(3) Further, if Assumption 4(iii') and

$$\|\Pi_J h_0 - \pi_J h_0\|_\infty \leq \text{const} \times \|h_0 - \Pi_J h_0\|_\infty \tag{2}$$

hold then Assumption 4(iii) is satisfied.

In light of Lemma 2.2 results (1) and (2), both Assumption 4(iii') and Condition (2) seem mild. In fact, Condition (2) is trivially satisfied when the basis for Ψ_J is a Riesz basis because then $\pi_J h_0 = \Pi_J h_0$ (see section 6 in Chen and Pouzo (2014)). See Lemma D.2 in the online Appendix D for more detailed relations among $\Pi_J h_0$, $\pi_J h_0$ and $Q_J h_0$.

Let $h_{0,J} \in \Psi_J$ solve $\inf_{h \in \Psi_J} \|h_0 - h\|_\infty$. Then:

$$\begin{aligned} \|h_0 - \Pi_J h_0\|_\infty &\leq \|h_0 - h_{0,J} + \Pi_J(h_0 - h_{0,J})\|_\infty \\ &\leq (1 + \|\Pi_J\|_\infty) \times \|h_0 - h_{0,J}\|_\infty \end{aligned}$$

where $\|\Pi_J\|_\infty$ is the Lebesgue constant for the sieve Ψ_J (see Lebesgue's lemma in DeVore and Lorentz (1993), page 30). Recently it has been established that $\|\Pi_J\|_\infty \lesssim 1$ when Ψ_J is spanned by a tensor product B-spline basis (Huang (2003)) or a tensor product CDV wavelet basis (Chen and Christensen (2014)). See DeVore and Lorentz (1993) and Belloni, Chernozhukov, Chetverikov, and Kato (2014) for examples of other bases with bounded Lebesgue constant or with Lebesgue constant diverging slowly with the sieve dimension.

The next lemma provides a bound on the sup-norm "bias" term.

Lemma 2.3 *Let Assumptions 1(iii), 3(ii) and 4 hold. Then:*
(1) $\|\widetilde{h} - \Pi_J h_0\|_\infty \leq O_p(1) \times \|h_0 - \Pi_J h_0\|_\infty$.
(2) $\|\widetilde{h} - h_0\|_\infty \leq O_p\left(1 + \|\Pi_J\|_\infty\right) \times \|h_0 - h_{0,J}\|_\infty$.

2.2.3 Sup-norm convergence rates

Lemmas 2.1(1) and 2.3 immediately yield the following general sup-norm rate result.

Theorem 2.1 *(1) Let Assumptions 1(i)(iii)(iv), 2(i)(ii), 3(ii)(iii), and 4 hold. Then:*

$$\|\widehat{h} - h_0\|_\infty = O_p\left(\|h_0 - \Pi_J h_0\|_\infty + \tau_J \xi_{\psi,J}\sqrt{(\log n)/(ne_J)}\right).$$

(2) Further, if the linear sieve Ψ_J satisfies $\|\Pi_J\|_\infty \lesssim 1$ and $\xi_{\psi,J}/\sqrt{e_J} \lesssim \sqrt{J}$, then

$$\|\widehat{h} - h_0\|_\infty = O_p\left(\|h_0 - h_{0,J}\|_\infty + \tau_J\sqrt{(J\log n)/n}\right).$$

The following corollary provides concrete sup-norm convergence rates of $\widehat{h}$ and its derivatives. To introduce the result, let $B^p_{\infty,\infty}$ denote the Hölder space of smoothness $p > 0$ and $\|\cdot\|_{B^p_{\infty,\infty}}$ denote its norm (see Triebel (2006)). Let $B_\infty(p, L) = \{h \in B^p_{\infty,\infty} : \|h\|_{B^p_{\infty,\infty}} \leq L\}$ denote a Hölder ball of smoothness $p > 0$ and radius $L \in (0, \infty)$. Let $\alpha_1, \ldots, \alpha_d$ be non-negative integers, let $|\alpha| = \alpha_1 + \ldots + \alpha_d$, and define

$$\partial^\alpha h(x) := \frac{\partial^{|\alpha|} h}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_d} x_d} h(x).$$

Of course, if $|\alpha| = 0$ then $\partial^\alpha h = h$.[6]

Corollary 2.1 *Let Assumptions 1(i)(ii)(iii) and 4 hold. Let $h_0 \in B_\infty(p, L)$, Ψ_J be spanned by a B-spline basis of order $\gamma > p$ or a CDV wavelet basis of regularity $\gamma > p$, B_K be spanned by a cosine, spline or wavelet basis.*
(1) If Assumption 3(ii) holds, then

$$\|\partial^\alpha \widetilde{h} - \partial^\alpha h_0\|_\infty = O_p\left(J^{-(p-|\alpha|)/d}\right) \quad \text{for all} \quad 0 \leq |\alpha| < p.$$

(2) Further if Assumptions 2(i)(ii) and 3(iii) hold, then

$$\|\partial^\alpha \widehat{h} - \partial^\alpha h_0\|_\infty = O_p\left(J^{-(p-|\alpha|)/d} + \tau_J J^{|\alpha|/d}\sqrt{(J\log n)/n}\right) \quad \text{for all} \quad 0 \leq |\alpha| < p.$$

(2.a) Mildly ill-posed case: with $p \geq d/2$ and $\delta \geq d/(p+\varsigma)$, choosing $J \asymp (n/\log n)^{d/(2(p+\varsigma)+d)}$ implies that Assumption 3(ii)(iii) holds and

$$\|\partial^\alpha \widehat{h} - \partial^\alpha h_0\|_\infty = O_p((n/\log n)^{-(p-|\alpha|)/(2(p+\varsigma)+d)}).$$

[6] If $|\alpha| > 0$ then we assume h and its derivatives can be continuously extended to an open set containing $\mathcal{X}$ so that $\partial^\alpha h(x)$ is well defined for all $x \in \mathcal{X}$.

(2.b) Severely ill-posed case: choosing $J = (c_0 \log n)^{d/\varsigma}$ *with* $c_0 \in (0,1)$ *implies that Assumption 3(ii)(iii) holds and*

$$\|\partial^\alpha \widehat{h} - \partial^\alpha h_0\|_\infty = O_p((\log n)^{-(p-|\alpha|)/\varsigma}) .$$

Corollary 2.1 is very useful for linearizing plug-in estimators of nonlinear functionals of h_0 to establish pointwise and uniform limit theory; see Appendices A and B. Corollary 2.1 is also useful for estimating functions with certain shape properties. For instance, if $h_0 : [a,b] \to \mathbb{R}$ is strictly monotone and/or strictly concave/convex then knowing that $\widehat{h}'(x)$ and/or $\widehat{h}''(x)$ converge uniformly to $h_0'(x)$ and/or $h_0''(x)$ implies that $\widehat{h}$ will also be strictly monotone and/or strictly concave/convex with probability approaching one.

2.3 Lower bounds on sup-norm rates

We now establish (minimax) optimality of the sup-norm rates obtained in Corollary 2.1. Previously, Hall and Horowitz (2005) and Chen and Reiss (2011) derived optimal L^2 norm rates for estimating h_0. We complement their analysis by deriving optimal sup-norm rates for estimating h_0 and its derivatives.

To establish a lower bound we require a link condition which measures how much the conditional expectation operator T smoothes out the structural function h_0. Consider the ball $\mathcal{H}_2(p,L) := \{h = \sum_{j,G,k} a_{j,k,G} \widetilde{\psi}_{j,k,G} : a_{j,k,G} \in \mathbb{R}, \sum_{j,G,k} 2^{jp} a_{j,k,G}^2 \leq L^2\}$ where $\widetilde{\psi}_{j,k,G}$ is a tensor product CDV wavelet basis for $[0,1]^d$ of regularity $\gamma > p > 0$ (see Appendix C for more details) and $L \in (0,\infty)$. The ball $\mathcal{H}_2(p,L)$ is equivalent to the Sobolev ball $B_2(p,L)$ (see Section 2.4) since for any $\mathcal{H}_2(p,L)$ there exists $L', L'' \in (0,\infty)$ such that $B_2(p,L') \subseteq \mathcal{H}_2(p,L) \subseteq B_2(p,L'')$.

Condition LB *(i) Assumption 1(i)–(iii) holds; (ii)* $E[u_i^2 | W_i = w] \geq \underline{\sigma}^2 > 0$ *uniformly for* $w \in \mathcal{W}$*; and (iii)* $\|Th\|_{L^2(W)}^2 \lesssim \sum_{j,G,k} \nu(2^j)^2 \langle h, \widetilde{\psi}_{j,k,G} \rangle_X^2$.

Condition LB(i)–(ii) is standard (see Hall and Horowitz (2005) and Chen and Reiss (2011)). Condition LB(iii) is a so-called link condition (Chen and Reiss, 2011). In an earlier version of the paper we derived a lower bound for h_0 in the mildly ill-posed case under the condition $\|Th\|_{L^2(X)}^2 \asymp \|h\|_{B_{2,2}^{-\varsigma}}$ for some $\varsigma > 0$, which corresponds to choosing $\nu(t) = t^{-\varsigma}$ in the above condition. Here we also allow for the severely ill-posed case, which corresponds to choosing $\nu(t) = \exp(-\frac{1}{2}t^\varsigma)$.

Let $\mathbb{P}_h$ denote the probability measure of the data when the structural function is h.

Theorem 2.2 *Let Condition LB hold for the NPIV model with a random sample* $\{(X_i, Y_i, W_i)\}_{i=1}^n$*. Then for any* $0 \leq |\alpha| < p$*:*

$$\liminf_{n\to\infty} \inf_{\widehat{g}_n} \sup_{h \in B_\infty(p,L)} \mathbb{P}_h \left(\|\widehat{g}_n - \partial^\alpha h\|_\infty \geq c(n/\log n)^{-(p-|\alpha|)/(2(p+\varsigma)+d)} \right) \geq c' > 0$$

in the mildly ill-posed case, and

$$\liminf_{n\to\infty}\inf_{\widehat{g}_n}\sup_{h\in B_\infty(p,L)}\mathbb{P}_h\left(\|\widehat{g}_n-\partial^\alpha h\|_\infty\geq c(\log n)^{-(p-|\alpha|)/\varsigma}\right)\geq c'>0$$

in the severely ill-posed case, where $\inf_{\widehat{g}_n}$ *denotes the infimum over all estimators of* $\partial^\alpha h$ *based on the sample of size* n*, and the finite positive constants* c, c' *do not depend on* n*.*

2.4 Optimal L^2-norm rates in derivative estimation

Here we show that the sieve NPIV estimator and its derivatives can attain the optimal L^2-norm convergence rates for estimating h_0 and its derivatives under much weaker conditions The optimal L^2-norm rates for sieve NPIV derivative estimation presented in this section are new, and should be very useful for inference on some nonlinear functionals involving derivatives such as $f(h) = \|\partial^\alpha h\|^2_{L^2(X)}$.

Theorem 2.3 *Let Assumptions 1(iii) and 4(i)(ii) hold and let* $\tau_J\zeta\sqrt{(\log J)/n}=o(1)$*. Then:*
(1) $\|\widetilde{h}-h_0\|_{L^2(X)}\leq O_p(1)\times\|h_0-\Pi_J h_0\|_{L^2(X)}$.
(2) Further, if Assumption 2(i) holds then

$$\|\widehat{h}-h_0\|_{L^2(X)}=O_p\left(\|h_0-\Pi_J h_0\|_{L^2(X)}+\tau_J\sqrt{J/n}\right).$$

The following corollary provides concrete L^2 norm convergence rates of $\widehat{h}$ and its derivatives. To introduce the result, let $\|\cdot\|_{B^p_{2,2}}$ denote the Sobolev norm of smoothness p (see Triebel (2006)), $B^p_{2,2}$ denote the Sobolev space of smoothness $p>0$, and $B_2(p,L)=\{h\in B^p_{2,2}:\|h\|_{B^p_{2,2}}\leq L\}$ denote a Sobolev ball of smoothness $p>0$ and radius $0<L<\infty$.

Corollary 2.2 *Let Assumptions 1(i)(ii)(iii) and 4(i)(ii) hold. Let* $h_0\in B_2(p,L)$*,* Ψ_J *be spanned by a cosine basis, B-spline basis of order* $\gamma>p$*, or CDV wavelet basis of regularity* $\gamma>p$*,* B_K *be spanned by a cosine, spline, or wavelet basis.*
(1) If $\tau_J\sqrt{(J\log J)/n}=o(1)$ *holds, then*

$$\|\partial^\alpha\widetilde{h}-\partial^\alpha h_0\|_{L^2(X)}=O_p\left(J^{-(p-|\alpha|)/d}\right)\quad \textit{for all}\quad 0\leq|\alpha|<p\,.$$

(2) Further if Assumption 2(i) holds, then

$$\|\partial^\alpha\widehat{h}-\partial^\alpha h_0\|_{L^2(X)}=O_p\left(J^{-(p-|\alpha|)/d}+\tau_J J^{|\alpha|/d}\sqrt{J/n}\right)\quad \textit{for all}\quad 0\leq|\alpha|<p\,.$$

(2.a) Mildly ill-posed case: choosing $J \asymp n^{d/(2(p+\varsigma)+d)}$ *yields* $\tau_J\sqrt{(J\log J)/n} = o(1)$ *and*

$$\|\partial^\alpha \widehat{h} - \partial^\alpha h_0\|_{L^2(X)} = O_p(n^{-(p-|\alpha|)/(2(p+\varsigma)+d)}).$$

(2.b) Severely ill-posed case: choosing $J = (c_0 \log n)^{d/\varsigma}$ *for any* $c_0 \in (0,1)$ *yields* $\tau_J\sqrt{(J\log J)/n} = o(1)$ *and*

$$\|\partial^\alpha \widehat{h} - \partial^\alpha h_0\|_{L^2(X)} = O_p((\log n)^{-(p-|\alpha|)/\varsigma}).$$

The conclusions of Corollary 2.2 hold if an arbitrary basis is used for B_K under the condition $\tau_J\zeta_b\sqrt{(\log J)/n} = o(1)$. Our next theorem shows that the rates obtained in Corollary 2.2 are in fact optimal. It extends the earlier work by Chen and Reiss (2011) on the minimax lower bound in L^2 loss for estimating h_0 to that for estimating the derivatives, allowing for both mildly and severely ill-posed NPIV problems.

Theorem 2.4 *Let Condition LB hold for the NPIV model with a random sample* $\{(X_i, Y_i, W_i)\}_{i=1}^n$. *Then for any* $0 \le |\alpha| < p$*:*

$$\liminf_{n\to\infty} \inf_{\widehat{g}_n} \sup_{h\in B_2(p,L)} \mathbb{P}_h\left(\|\widehat{g}_n - \partial^\alpha h\|_{L^2(X)} \ge cn^{-(p-|\alpha|)/(2(p+\varsigma)+d)}\right) \ge c' > 0$$

in the mildly ill-posed case, and

$$\liminf_{n\to\infty} \inf_{\widehat{g}_n} \sup_{h\in B_2(p,L)} \mathbb{P}_h\left(\|\widehat{g}_n - \partial^\alpha h\|_{L^2(X)} \ge c(\log n)^{-(p-|\alpha|)/\varsigma}\right) \ge c' > 0$$

in the severely ill-posed case, where $\inf_{\widehat{g}_n}$ *denotes the infimum over all estimators of* $\partial^\alpha h$ *based on the sample of size* n*, and the finite positive constants* c, c' *do not depend on* n*.*

According to Theorems 2.2 and 2.4, the minimax lower bounds in sup-norm loss for estimating h_0 and its derivatives coincide with those in L^2 loss for severely ill-posed NPIV problems, and are only a factor of $[\log(n)]^\epsilon$ (with $\epsilon = \frac{p-|\alpha|}{2(p+\varsigma)+d} < \frac{p}{2p+d} < \frac{1}{2}$) worse than those in L^2 loss for mildly ill-posed problems.

2.5 Models with endogenous and exogenous regressors

We finish by discussing briefly models of the form

$$Y_i = h_0(X_{1i}, Z_i) + u_i \tag{3}$$

where X_{1i} is a vector of endogenous regressors and Z_i is a vector of exogenous regressors. Let $X_i = (X_{1i}', Z_i')'$. Here the vector of instrumental variables W_i is of the form $W_i = (W_{1i}', Z_i')'$ where

W_{1i} are instruments for X_{1i}. We refer to this as the "partially endogenous case".

The sieve NPIV estimator is implemented in exactly the same way as the "fully endogenous" setting in which X_i consists only of endogenous variables, just as with 2SLS estimation with endogenous and exogenous variables (Newey and Powell, 2003; Ai and Chen, 2003; Blundell et al., 2007). Other NPIV estimators based on first estimating the conditional densities of the regressors variables and instrumental variables must be implemented separately at each value of z in the partially endogenous case (Hall and Horowitz, 2005; Horowitz, 2011; Gagliardini and Scaillet, 2012).

Our convergence rates presented in Sections 2.2 and 2.4 apply equally to the partially endogenous model (3) under the stated regularity conditions: all that differs between the two cases is the interpretation of the sieve measure of ill-posedness.

Consider first the fully endogenous case where $T : L^2(X) \to L^2(W)$ is compact. Then T admits a singular value decomposition (SVD) $\{\phi_{0j}, \phi_{1j}, \mu_j\}_{j=1}^{\infty}$ where $(T^*T)^{1/2}\phi_{0j} = \mu_j\phi_{0j}$, $\mu_j \geq \mu_{j+1}$ for each j and $\{\phi_{0j}\}_{j=1}^{\infty}$ and $\{\phi_{1j}\}_{j=1}^{\infty}$ are orthonormal bases for $L^2(X)$ and $L^2(W)$, respectively. Suppose that Ψ_J spans $\phi_{0j}, \ldots, \phi_{0J}$. Then the sieve measure of ill-posedness is $\tau_J = \mu_J^{-1}$ (see Blundell et al. (2007)). Now consider the partially endogenous case. Similar to Horowitz (2011), we suppose that for each value of z the conditional expectation operator $T_z : L^2(X_1|Z = z) \to L^2(W_1|Z = z)$ given by $(T_z h)(w_1) = E[h(X_1)|W_{1i} = w_1, Z_i = z]$ is compact. Then each T_z admits a SVD $\{\phi_{0j,z}, \phi_{1j,z}, \mu_{j,z}\}_{j=1}^{\infty}$ where $T_z\phi_{0j,z} = \mu_{j,z}\phi_{1j,z}$, $(T_z^*T_z)^{1/2}\phi_{0j,z} = \mu_{j,z}\phi_{0j,z}$, $(T_z T_z^*)^{1/2}\phi_{1j,z} = \mu_{j,z}\phi_{1j,z}$, $\mu_{j,z} \geq \mu_{j+1,z}$ for each j and z, and $\{\phi_{0j,z}\}_{j=1}^{\infty}$ and $\{\phi_{1j,z}\}_{j=1}^{\infty}$ are orthonormal bases for $L^2(X_1|Z = z)$ and $L^2(W_1|Z = z)$, respectively, for each z. The following result adapts Lemma 1 of Blundell et al. (2007) to the partially endogenous setting.

Lemma 2.4 *Let T_z be compact with SVD $\{\phi_{0j,z}, \phi_{1j,z}, \mu_{j,z}\}_{j=1}^{\infty}$ for each z. Let $\mu_j^2 = E[\mu_{j,Z_i}^2]$ and $\phi_{0j}(\cdot, z) = \phi_{0j,z}(\cdot)$ for each z and j. Then: (1) $\tau_J \geq \mu_J^{-1}$.*
(2) If, in addition, $\phi_{01}, \ldots, \phi_{0J} \in \Psi_J$, then: $\tau_J \leq \mu_J^{-1}$.

The following stylized example illustrates the behavior in the partially endogenous case relative to that in the fully endogenous case. Let X_{1i}, W_{1i} and Z_i be scalar random variables and let $(X_{1i}, W_{1i}, Z_i)'$ be distributed as

$$\begin{pmatrix} X_{1i} \\ W_{1i} \\ Z_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{XW} & \rho_{XZ} \\ \rho_{XW} & 1 & \rho_{WZ} \\ \rho_{XZ} & \rho_{WZ} & 1 \end{pmatrix}\right)$$

and $\rho_{XW}, \rho_{XZ}, \rho_{WZ}$ are such that the covariance matrix is positive definite. Then

$$\begin{pmatrix} \frac{X_{1i}-\rho_{XZ}z}{\sqrt{1-\rho_{XZ}^2}} \\ \frac{W_{1i}-\rho_{WZ}z}{\sqrt{1-\rho_{WZ}^2}} \end{pmatrix} \Bigg| Z_i = z \Bigg) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{XW|Z} \\ \rho_{XW|Z} & 1 \end{pmatrix}\right) \tag{4}$$

where

$$\rho_{XW|Z} = \frac{\rho_{XW} - \rho_{XZ}\rho_{WZ}}{\sqrt{(1-\rho_{XZ}^2)(1-\rho_{WZ}^2)}}$$

is the partial correlation between X_{1i} and W_{1i} given Z_i.

For each $j \geq 1$ let H_j denote the jth Hermite polynomial subject to the normalizations $H_0(x) = 1$ and $\int_{-\infty}^{\infty} H_j(x)H_k(x)\,\mathrm{d}\Phi(x) = \delta_{jk}$ for $0 \leq j, k < \infty$ where δ_{jk} denotes the Kronecker delta and Φ is the standard normal distribution. Since $T_z : L^2(X_1|Z=z) \to L^2(W_1|Z=z)$ is compact for each z, it follows from Mehler's formula that each T_z has a SVD $\{\phi_{0j,z}, \phi_{1j,z}, \mu_{j,z}\}_{j=1}^{\infty}$ given by

$$\phi_{0j,z}(x_1) = H_{j-1}\left(\frac{x_1 - \rho_{XZ}z}{\sqrt{1-\rho_{XZ}^2}}\right), \quad \phi_{1j,z}(w_1) = H_{j-1}\left(\frac{w_1 - \rho_{WZ}z}{\sqrt{1-\rho_{WZ}^2}}\right), \quad \mu_{j,z} = |\rho_{XW|Z}|^{j-1}.$$

Since $\mu_{J,z} = |\rho_{XW|Z}|^{J-1}$ for each z, we have $\mu_J = |\rho_{XW|Z}|^{J-1} \asymp |\rho_{XW|Z}|^J$. If X_{1i} and W_{1i} are uncorrelated with Z_i then $\mu_J = |\rho|^{J-1}$ where $\rho = \rho_{XW}$.

Now compare the partially endogenous case just described with the following fully-endogenous model in which X_i and W_i are bivariate with

$$\begin{pmatrix} X_{1i} \\ X_{2i} \\ W_{1i} \\ W_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \rho_1 & 0 \\ 0 & 1 & 0 & \rho_2 \\ \rho_1 & 0 & 1 & 0 \\ 0 & \rho_2 & 0 & 1 \end{pmatrix}\right)$$

where ρ_1 and ρ_2 are such that the covariance matrix is invertible. It is straightforward to verify that T has singular value decomposition with

$$\phi_{0j}(x) = H_{j-1}(x_1)H_{j-1}(x_2) \quad \phi_{1j}(w) = H_{j-1}(w_1)H_{j-2}(w_2), \quad \mu_j = |\rho_1\rho_2|^{j-1}.$$

In particular, when $\rho_1 = \rho_2 = \rho$ we have $\mu_J = \rho^{2(J-1)}$.

In both of the preceding examples, h_0 is a function of two random variables (X_1, Z). The degree of ill-posedness in the partially endogenous case is $|\rho|^{-(J-1)}$ where ρ is the correlation between the endogenous regressor and its instrument. The degree of ill-posedness increases to $(\rho^2)^{-(J-1)}$ in the fully endogenous case when each endogenous regressor has correlation ρ with its instrument.

3 Adaptive estimation in sup-norm loss

We now propose a simple, data-driven method for choosing the sieve dimension, which is a novel extension of the balancing principle of Lepskii (1990) to nonparametric models with endogeneity.

Our selection criterion is optimal in that the resulting sieve NPIV estimator of h_0 and its derivatives attain the optimal sup-norm rates.

To describe our data-driven method for choosing J, let $J_{\min} = \lfloor \log\log n \rfloor$ and let $\underline{J}_{\max}$ and $\bar{J}_{\max}$ be increasing sequences of integers which index the sieve dimension Ψ_J where $J_{\min} < \underline{J}_{\max} \leq \bar{J}_{\max}$.[7] Let $\mathbb{N}^*$ denote the sequence of integers which index the dimension of the sieve spaces Ψ_J and let $\underline{I}_J = \{j \in \mathbb{N}^* : J_{\min} \leq j \leq \underline{J}_{\max}\}$ and $\bar{I}_J = \{j \in \mathbb{N}^* : J_{\min} \leq j \leq \bar{J}_{\max}\}$. Finally, let $\widehat{J}_{\max}$ be a possibly random integer such that $\underline{J}_{\max} \leq \widehat{J}_{\max} \leq \bar{J}_{\max}$ wpa1 (we introduce such a data-driven choice below) and let $\widehat{I}_J = \{j \in \mathbb{N}^* : J_{\min} \leq j \leq \widehat{J}_{\max}\}$.

The oracle and data-driven index sets are defined as

$$\begin{aligned}
\mathcal{J}_0 &= \left\{ j \in \underline{I}_J : \|h_0 - \Pi_j h_0\|_\infty \leq C_0 V_{\sup}(j) \right\} \\
\widehat{\mathcal{J}} &= \left\{ j \in \widehat{I}_J : \|\widehat{h}_j - \widehat{h}_l\|_\infty \leq \sqrt{2}\overline{\sigma}(\widehat{V}_{\sup}(j) + \widehat{V}_{\sup}(l)) \text{ for all } l \in \widehat{I}_J \text{ with } l \geq j \right\}
\end{aligned}$$

respectively, where C_0 is a finite positive constant and

$$\begin{aligned}
V_{\sup}(j) &= \tau_j \xi_{\psi,j} \sqrt{(\log n)/(n e_j)} \\
\widehat{V}_{\sup}(j) &= \widehat{\tau}_j \xi_{\psi,j} \sqrt{(\log n)/(n \widehat{e}_j)}
\end{aligned}$$

where $\widehat{e}_j = \lambda_{\min}(\widehat{G}_{\psi,j})$ with $\widehat{G}_{\psi,j} = \Psi'\Psi/n$, and $\widehat{\tau}_j$ is an estimator of the degree of ill-posedness τ_j:

$$\widehat{\tau}_j = \sup_{h \in \Psi_j : h \neq 0} \sqrt{\frac{\frac{1}{n}\sum_{i=1}^n h(X_i)^2}{\frac{1}{n}\sum_{i=1}^n \widehat{E}[h(X_i)|W_i]^2}}$$

where

$$\widehat{E}[h(X_i)|W_i = w] = b^K(w)(B'B/n)^- \left(\frac{1}{n}\sum_{i=1}^n b^K(W_i) h(X_i) \right).$$

is a series regression estimator of $E[h(X_i)|W_i = w]$. The variational characterization of singular values gives an alternative, computationally simple expression for $\widehat{\tau}_j$, namely:

$$\widehat{\tau}_j = \frac{1}{s_{\min}\left((B'B/n)^{-1/2} (B'\Psi/n)(\Psi'\Psi/n)^{-1/2} \right)} \tag{5}$$

where $s_{\min}(A)$ denotes the smallest singular value of the matrix A and $^{-1/2}$ denotes the inverse of the positive definite square root. We specify the smoothing parameter K as a known function of the regularization parameter J according to a rule $K : \mathbb{N} \to \mathbb{N}$ for which $j \leq K(j) \leq C_K j$ for all $j \in \mathbb{N}$ and for some $1 \leq C_K < \infty$, such as $K(j) = j$ or $K(j) = 2j$. In what follows we let $\widehat{h}_J$ denote the sieve NPIV estimator with regularization parameter J and smoothing parameter $K = K(J)$.

[7] This choice of $J_{\min}$ ensures $J_{\min}$ grows slower than the optimal choice of J in the mildly and severely ill-posed cases. For NPIV models in which τ_J grows faster than the severely ill-posed case, we can take $J_{\min}$ to be an even more slowly growing function of n than $\lfloor \log\log n \rfloor$.

The set $\mathcal{J}_0$ is the unobservable *oracle set*: it depends on the unknown smoothness of h_0 and the unknown degree of ill-posedness. We refer to

$$J_0 = \min_{j \in \mathcal{J}_0} j$$

as the *oracle* choice of J as it balances the bias and variance terms asymptotically. As a consequence, we will refer to $\widehat{h}_{J_0}$ as the *oracle estimator*.

The set $\widehat{\mathcal{J}}$ is straightforward to construct from the data as it depends entirely on observables and $\overline{\sigma}$, about which the researcher may have a priori information.[8] Our *data-driven* choice of J is

$$\widehat{J} = \min_{j \in \widehat{\mathcal{J}}} j$$

and $\widehat{h}_{\widehat{J}}$ is our data-driven sieve NPIV estimator. Below we will establish an oracle inequality which shows that, with high probability, the sup-norm loss of the data-driven estimator $\widehat{h}_{\widehat{J}}$ is bounded by a multiple of the sup-norm loss of the oracle estimator $\widehat{h}_{J_0}$.

Our adaptive procedure relies on balancing the bias and variance terms asymptotically, so we need to bound the variance term up to known or estimable constants, which explains why we use the variance bound in Lemma 2.1(1), i.e., $V_{\sup}(j) = \tau_j \xi_{\psi,j} \sqrt{(\log n)/(n e_j)}$. This variance bound ensures that $\|\widehat{h}_j - \widetilde{h}_j\|_\infty \leq \sqrt{2}\overline{\sigma} V_{\sup}(j)$ holds uniformly over a range of j wpa1, and does not affect the attainability of the optimal sup-norm rates using spline or wavelet bases for Ψ_J (see Corollary 2.1).

3.1 Sup-norm rate-adaptivity to the oracle

In this section we establish oracle properties of our data-driven estimator $\widehat{h}_{\widehat{J}}$.

Let $\kappa_b(K)$ denote the condition number of $G_{b,K}$ and let $\kappa_\psi(j)$ denote the condition number of $G_{\psi,j}$. To simplify the notation in the following presentation, we assume for simplicity that $\zeta_{b,K(j)}$, $\zeta_{\psi,j}$, τ_j, e_j^{-1}, $\kappa_b(K(j))$, $\kappa_\psi(j)$ are all (weakly) increasing on $\bar{I}_J$. Our proofs and conditions can easily be adapted to dispense with this assumption at the cost of more complicated notation. Let $\overline{K}_{\max} = K(\bar{J}_{\max})$ and $\bar{\zeta} = \zeta(\bar{J}_{\max}) = \zeta_{b,\overline{K}_{\max}} \vee \zeta_{\psi,\bar{J}_{\max}}$.

Assumption 3 (continued) *(iv)* $\tau_{\bar{J}_{\max}} \bar{\zeta}^2 \sqrt{(\log n)/n} = o(1)$ *and* $\bar{J}_{\max}^{2+\epsilon}/n = O(1)$ *for some* $\epsilon > 0$*; (v)* $\bar{\zeta}^{(2+\delta)/\delta} \sqrt{(\log n)/n} = o(1)$*; (vi)* $\kappa_b(K) = O(\zeta_{b,K})$ *and* $\kappa_\psi(J) = O(\zeta_{\psi,J})$*.*

Assumption 3(iv)(v) are uniform (for $j \in \bar{I}_J$) versions of Assumption 3(ii)(iii). The second part of Assumption 3(iv) may be replaced be an "enough ill-posedness" condition requiring τ_J to grow

[8] Our procedure remains valid whenever $\overline{\sigma}$ in the definition of $\widehat{\mathcal{J}}$ is replaced by a consistent estimator. Further, we show in the Monte Carlo exercise below that the procedure is reasonably robust to choosing too small a value of $\overline{\sigma}$.

faster than J^α for some $\alpha > 0$ (cf. Assumption 6(iii) of Horowitz (2014)). Assumption 3(vi) is a further condition on $J_{\max}$ which allows for consistent estimation of τ_j and e_j for all $j \in I_J$. We show in Appendix C that $\kappa_b(K) = O(1)$ and $\kappa_\psi(J) = O(1)$ when B_K and Ψ_J are spanned by (tensor-product) spline or wavelet bases, in which case Assumption 3(vi) holds trivially.

Theorem 3.1 *Let Assumptions 1, 2(i)(ii), 3(iv)(vi), and 4 hold and let* $\underline{J}_{\max} \leq \widehat{J}_{\max} \leq \overline{J}_{\max}$ *hold wpa1. Then:*

$$\|\widehat{h}_{\widehat{J}} - h_0\|_\infty \leq \|\widehat{h}_{J_0} - h_0\|_\infty + 3\overline{\sigma} V_{\sup}(J_0)$$

holds wpa1, and so

$$\|\widehat{h}_{\widehat{J}} - h_0\|_\infty = O_p\Big(\|h_0 - \Pi_{J_0} h_0\|_\infty + \tau_{J_0}\xi_{\psi,J_0}\sqrt{(\log n)/(ne_{J_0})}\Big)\,.$$

Corollary 3.1 *Let the assumptions of Theorem 3.1 hold with* $h_0 \in B_\infty(p, L)$ *and let* Ψ_J *be spanned by a (tensor product) B-spline or CDV wavelet basis. Then:*

$$\|\partial^\alpha \widehat{h}_{\widehat{J}} - \partial^\alpha h_0\|_\infty = O_p\Big(J_0^{-(p-|\alpha|)/d} + \tau_{J_0} J_0^{|\alpha|/d}\sqrt{(J_0 \log n)/n}\Big) \quad \textit{for all} \quad 0 \leq |\alpha| < p\,.$$

(1) Mildly ill-posed case: if $p > d/2$ *and* $\delta > d/(p+\varsigma)$ *then for all* $0 \leq |\alpha| < p$,

$$\|\partial^\alpha \widehat{h}_{\widehat{J}} - \partial^\alpha h_0\|_\infty = O_p((n/\log n)^{-(p-|\alpha|)/(2(p+\varsigma)+d)})\,.$$

(2) Severely ill-posed case: for all $0 \leq |\alpha| < p$,

$$\|\partial^\alpha \widehat{h}_{\widehat{J}} - \partial^\alpha h_0\|_\infty = O_p((\log n)^{-(p-|\alpha|)/\varsigma})\,.$$

Previously, Horowitz (2014) introduced a model selection procedure to choose J for his modified orthogonal series NPIV estimator (i.e., a series 2SLS estimator with $K(J) = J$, $b^K = \psi^J$ being orthonormal basis in $L^2([0,1])$), and showed that his data-driven choice leads to near L^2-norm rate adaptivity in that his estimator is a factor of $\sqrt{\log n}$ slower than the optimal L^2 norm convergence rate for estimating h_0 (see Theorem 3.2 of Horowitz (2014)).[9] It follows from Corollary 3.1 with $|\alpha| = 0$ that our data-driven estimator $\widehat{h}_{\widehat{J}}$ converges in sup norm (and therefore in L^2 norm) faster than that of Horowitz (2014))'s in L^2 norm.

Recently, Breunig and Johannes (2013) also applied Lepski's method to study near L^2-norm adaptive estimation of linear functionals of NPIV models.[10] Gautier and LePennec (2011) proposed a data-driven method for choosing the regularization parameter in a random coefficient binary choice

[9] See Loubes and Marteau (2012) and Johannes and Schwarz (2013) for near L^2-norm rate adaptivity of estimators similar to Horowitz (2014)'s when the eigenfunctions of the conditional expectation operator are known.

[10] Lepski methods have been used elsewhere in econometrics. See, e.g., Andrews and Sun (2004) for adaptive estimation of the long memory parameter.

model that is sup-norm rate adaptive for a mildly ill-posed deconvolution type problem. See Hoffmann and Reiss (2008) and references therein for other L^2-norm rate-adaptive schemes for ill-posed inverse problems in which the operator is known up to a random perturbation but is not estimated from the data. In work that is concurrent with ours, Liu and Tao (2014) show that the model selection approach of Li (1987) may be used to choose the sieve dimension to minimize empirical MSE in NPIV models with known homoskedastic errors.[11] Our procedure appears to be the first to attain both sup-norm and L^2-norm rate-adaptive estimation of h_0 and its derivatives for severely ill-posed NPIV models, and sup-norm rate-adaptive and near L^2-norm rate adaptive for mildly ill-posed NPIV models.

3.2 A data-driven upper bound for the index set

Theorem 3.1 is valid for an arbitrary estimator $\widehat{J}_{\max}$ of the upper level of the index set $\widehat{\mathcal{J}}$. We now propose such an estimator and show that it verifies the conditions of Theorem 3.1.

We propose choosing the maximum $\widehat{J}_{\max}$ of $\widehat{\mathcal{J}}$ using the estimator

$$\widehat{J}_{\max} := \min\left\{ J > J_{\min} : \widehat{\tau}_J [\zeta(J)]^2 \sqrt{L(J)(\log n)/n} \geq 1 \right\} \tag{6}$$

where $L(J) = a \log\log J$ for some positive constant a and we take $[\zeta(J)]^2 = J$ if B_K and Ψ_J are spanned by a spline, wavelet, or cosine basis, and $[\zeta(J)]^2 = J^2$ if B_K and Ψ_J are spanned by orthogonal polynomial basis. The following result shows that $\widehat{J}_{\max}$ defined in (6) satisfies the conditions of Theorem 3.1.

Theorem 3.2 *Let Assumptions 1(iii), 3(vi) and 4(i) hold. Then there exists deterministic sequences of integers $\underline{J}_{\max}, \bar{J}_{\max} \nearrow \infty$ such that $\tau_{\bar{J}_{\max}} \bar{\zeta}^2 \sqrt{(\log n)/n} = o(1)$ and $\underline{J}_{\max} \leq \widehat{J}_{\max} \leq \bar{J}_{\max}$ holds wpa1.*

3.3 Monte Carlo

In this section we evaluate the performance of our adaptive procedure. We use the experimental design of Newey and Powell (2003), in which IID draws are generated from

$$\begin{pmatrix} U_i \\ V_i^* \\ W_i^* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right)$$

[11] See Centorrino (2014) and Sueishi (2012) for data-driven choice of regularization parameters based on minimizing reduced-form empirical MSE. These papers do not study whether or not their procedures might lead to optimal or near-optimal convergence rates for estimating h_0.

from which we then set $X_i^* = W_i^* + V_i^*$. To ensure compact support of the regressor and instrument, we rescale X_i^* and W_i^* by defining $X_i = \Phi(X_i^*/\sqrt{2})$ and $W_i = \Phi(W_i^*)$ where Φ is the Gaussian cdf. We use $h_0(x) = 4x - 2$ for our *linear* design and $h_0(x) = \log(|6x-3|+1)\text{sgn}(x-\frac{1}{2})$ for our *nonlinear* design (our nonlinear h_0 is a re-scaled version of the h_0 used in Newey and Powell (2003)). Note that the nonlinear h_0 belongs to a Hölder ball of smoothness p with $1 < p < 2$.

For both designs, we generate 1000 samples of length $n = 1000$ and $n = 5000$ and implement our procedure using cubic and quartic B-spline bases with interior knots placed evenly. We use $0, 1, 3, 7, 15, \ldots$ interior knots so that the sieve spaces B_K and Ψ_J are nested as K, J increase. We then repeat the experiments using Legendre polynomial bases (orthonormalized with respect to the $L^2([0,1])$ inner product). Note that the results using a Legendre polynomial basis for the space Ψ_J are incompatible with our earlier theory on attainability of optimal sup-norm rates. We set $\overline{\sigma} = 1$ in $\widehat{\mathcal{J}}$ for all simulation designs and then repeat the experiments with $\overline{\sigma} = 0.1$ to investigate the sensitivity of our estimator to the user-specified value $\overline{\sigma}$ ($\overline{\sigma} = 1$ is the correct conditional variance of u). We choose $\widehat{J}_{\max}$ as described in Section 3.2 with $a = \frac{1}{10}$ (the results were insensitive to the choice of a). For each sample we calculate the sup-norm and L^2-norm loss of our estimator $\widehat{h}_{\widehat{J}}$ and the sup-norm relative error ratio

$$\frac{\|\widehat{h}_{\widehat{J}} - h_0\|_\infty}{\|\widehat{h}_{J_\infty} - h_0\|_\infty} \quad \text{where} \quad J_\infty = \text{argmin}_{j \in I_J} \|\widehat{h}_j - h_0\|_\infty \tag{7}$$

and the L^2-norm relative error ratio

$$\frac{\|\widehat{h}_{\widehat{J}} - h_0\|_{L^2(X)}}{\|\widehat{h}_{J_2} - h_0\|_{L^2(X)}} \quad \text{where} \quad J_2 = \text{argmin}_{j \in I_J} \|\widehat{h}_j - h_0\|_{L^2(X)} \tag{8}$$

where J_∞ and J_2 are the (infeasible) choices of J which minimize the sup and L^2-norm errors of h_J in the sample. Finally, we take the average of each of $\|\widehat{h}_{\widehat{J}} - h_0\|_\infty$, $\|\widehat{h}_{\widehat{J}} - h_0\|_{L^2(X)}$, and equations (7) and (8) across the 1000 samples.

The results of the MC exercise with $n = 1000$ are presented in Tables 1 and 2 and may be summarized as follows:
(1) When implemented with a B-spline basis for Ψ_J, the sup-norm loss of our data-driven estimator $\widehat{h}_{\widehat{J}}$ is at most 11% larger than that of the infeasible estimator which minimizes the sup-norm loss in each sample. Further, the L^2-norm loss of $\widehat{h}_{\widehat{J}}$ is at most 6% larger than that of the infeasible estimator which minimizes the L^2-norm loss in each sample.
(2) Reducing $\overline{\sigma}$ from 1 (correct) to 0.1 (incorrect) has little, if any, effect on the performance of $\widehat{h}_{\widehat{J}}$ when a B-spline basis is used for Ψ_J.
(3) The data-driven estimator $\widehat{h}_{\widehat{J}}$ again performs well with Legendre polynomial bases and $\overline{\sigma} = 1$, with sup-norm loss at most 1% larger than that of the infeasible choice of J for the linear design, and at most 15% larger than that of the infeasible choice in the nonlinear design. Similar results

		$K(J) = J$				$K(J) = 2J$			
		$\overline{\sigma} = 1$		$\overline{\sigma} = 0.1$		$\overline{\sigma} = 1$		$\overline{\sigma} = 0.1$	
r_J	r_K	sup	L^2	sup	L^2	sup	L^2	sup	L^2
		Design 1: Linear h_0							
4	4	1.0287	1.0003	1.0326	1.0030	1.0874	1.0383	1.0994	1.0530
4	5	1.0696	1.0293	1.0835	1.0447	1.0579	1.0196	1.0879	1.0456
5	5	1.0712	1.0198	1.0712	1.0198	1.1092	1.0560	1.1092	1.0560
4	Leg	1.0332	1.0005	1.0385	1.0067	1.0469	1.0106	1.1004	1.0569
5	Leg	1.0745	1.0208	1.0745	1.0208	1.0558	1.0278	1.0558	1.0278
Leg	Leg	1.0150	1.0000	4.6014	3.1509	1.0175	1.0031	6.2664	4.1851
		Design 2: Nonlinear h_0							
4	4	1.0235	1.0006	1.0278	1.0032	1.0782	1.0325	1.0885	1.0479
4	5	1.0623	1.0266	1.0691	1.0346	1.0486	1.0167	1.0804	1.0401
5	5	1.0740	1.0216	1.0740	1.0216	1.1138	1.0605	1.1138	1.0605
4	Leg	1.0280	1.0016	1.0336	1.0078	1.0406	1.0104	1.0914	1.0520
5	Leg	1.0785	1.0231	1.0785	1.0231	1.0613	1.0355	1.0613	1.0355
Leg	Leg	1.1185	1.1516	1.8019	1.4263	1.1418	1.1883	1.7814	1.4307

Table 1: Average sup-norm and L^2-norm relative error ratios (see equations (7) and (8)) across MC simulations with $n = 1000$. Results are presented for B-spline and Legendre polynomial bases with two different rules for $K(J)$. Results for $r_J = 4$ ($r_J = 5$) use a cubic (quartic) B-spline basis for Ψ_J, $r_J =$ Leg use a Legendre polynomial basis for Ψ_J. The r_K column specifies the basis for B_K similarly. Columns headed $\overline{\sigma} = 1$ and $\overline{\sigma} = 0.1$ correspond to implementing $\widehat{h}_{\widehat{J}}$ with the correct and incorrect value of $\overline{\sigma}$, respectively.

		$K(J) = J$				$K(J) = 2J$			
		$\overline{\sigma} = 1$		$\overline{\sigma} = 0.1$		$\overline{\sigma} = 1$		$\overline{\sigma} = 0.1$	
r_J	r_K	sup	L^2	sup	L^2	sup	L^2	sup	L^2
		Design 1: Linear h_0							
4	4	0.4262	0.1547	0.4298	0.1556	0.4188	0.1526	0.4233	0.1548
4	5	0.4179	0.1524	0.4226	0.1545	0.3918	0.1439	0.4038	0.1483
5	5	0.6633	0.2355	0.6633	0.2355	0.6366	0.2277	0.6366	0.2277
4	Leg	0.4262	0.1547	0.4312	0.1566	0.3778	0.1388	0.3962	0.1452
5	Leg	0.6633	0.2355	0.6633	0.2355	0.5977	0.2155	0.5977	0.2155
		Design 2: Nonlinear h_0							
4	4	0.4343	0.1621	0.4380	0.1631	0.4271	0.1601	0.4324	0.1628
4	5	0.4262	0.1600	0.4290	0.1613	0.4002	0.1518	0.4123	0.1560
5	5	0.6726	0.2407	0.6726	0.2407	0.6471	0.2330	0.6471	0.2330
4	Leg	0.4343	0.1621	0.4394	0.1640	0.3854	0.1475	0.4030	0.1534
5	Leg	0.6726	0.2407	0.6726	0.2407	0.6068	0.2215	0.6068	0.2215

Table 2: Average sup-norm error and L^2-norm error of $\widehat{h}_{\widehat{J}}$ across MC simulations with $n = 1000$. Columns and row headings are as described in Table 1.

		$K(J)=J$				$K(J)=2J$			
		$\overline{\sigma}=1$		$\overline{\sigma}=0.1$		$\overline{\sigma}=1$		$\overline{\sigma}=0.1$	
r_J	r_K	sup	L^2	sup	L^2	sup	L^2	sup	L^2
		Design 1: Linear h_0							
4	4	1.0239	1.0004	1.0239	1.0004	1.0638	1.0345	1.0653	1.0358
4	5	1.0533	1.0274	1.0533	1.0274	1.0415	1.0091	1.0444	1.0120
5	5	1.0616	1.0048	1.0616	1.0048	1.0987	1.0418	1.0987	1.0418
4	Leg	1.0244	1.0004	1.0244	1.0004	1.0339	1.0094	1.0339	1.0094
5	Leg	1.0635	1.0043	1.0635	1.0043	1.0467	1.0143	1.0467	1.0143
Leg	Leg	1.0136	1.0000	4.3937	3.0114	1.0135	1.0010	3.9764	2.7687
		Design 2: Nonlinear h_0							
4	4	1.0168	1.0027	1.0168	1.0027	1.0435	1.0231	1.0448	1.0244
4	5	1.0377	1.0200	1.0377	1.0200	1.0233	1.0092	1.0258	1.0123
5	5	1.0715	1.0175	1.0715	1.0175	1.0967	1.0508	1.0967	1.0508
4	Leg	1.0181	1.0028	1.0181	1.0028	1.0192	1.0091	1.0192	1.0091
5	Leg	1.0743	1.0176	1.0743	1.0176	1.0588	1.0386	1.0588	1.0386
Leg	Leg	1.3855	1.6588	1.5866	1.4010	1.4246	1.7316	1.4740	1.3321

Table 3: Average sup-norm and L^2-norm relative error ratios (see equations (7) and (8)) across MC simulations with $n = 5000$. Columns and row headings are as described in Table 1.

		$K(J)=J$				$K(J)=2J$			
		$\overline{\sigma}=1$		$\overline{\sigma}=0.1$		$\overline{\sigma}=1$		$\overline{\sigma}=0.1$	
r_J	r_K	sup	L^2	sup	L^2	sup	L^2	sup	L^2
		Design 1: Linear h_0							
4	4	0.1854	0.0669	0.1854	0.0669	0.1851	0.0668	0.1857	0.0669
4	5	0.1851	0.0667	0.1851	0.0667	0.1735	0.0626	0.1742	0.0628
5	5	0.3012	0.1051	0.3012	0.1051	0.2893	0.1018	0.2893	0.1018
4	Leg	0.1854	0.0669	0.1854	0.0669	0.1703	0.0616	0.1703	0.0616
5	Leg	0.3012	0.1051	0.3012	0.1051	0.2684	0.0965	0.2684	0.0965
		Design 2: Nonlinear h_0							
4	4	0.2037	0.0822	0.2037	0.0822	0.2031	0.0820	0.2037	0.0822
4	5	0.2031	0.0820	0.2031	0.0820	0.1921	0.0786	0.1928	0.0789
5	5	0.3150	0.1157	0.3150	0.1157	0.3044	0.1128	0.3044	0.1128
4	Leg	0.2037	0.0822	0.2037	0.0822	0.1889	0.0779	0.1889	0.0779
5	Leg	0.3150	0.1157	0.3150	0.1157	0.2844	0.1082	0.2844	0.1082

Table 4: Average sup-norm error and L^2-norm error of $\widehat{h}_{\widehat{J}}$ across MC simulations with $n = 5000$. Columns and row headings are as described in Table 1.

are obtained for L^2 loss.
(4) With a Legendre basis, the estimator appears to perform considerably worse with the (incorrect) $\overline{\sigma} = 0.1$, especially in the linear design. This merits an explanation. In the linear case, the true model is obtained with $J = 2$. As such, the L^2 and sup-norm error of the infeasible estimators are very small. When $\overline{\sigma} = 0.1$ the Lepski procedure is less conservative and the data-driven estimator $\widehat{h}_{\widehat{J}}$ is slightly more noisy. This noise is amplified in the relative error ratios because the loss for the infeasible estimators in this design is very small.
(5) The relative error of the estimators with $K > J$ is similar to than that obtained with $K = J$. The absolute error (see Table 2) of the estimators with $K > J$ was slightly better than that with $K = J$. This emphasizes that the critical smoothing parameter is J; the choice of K is of higher-order importance.

Tables 3 and 4 display the results for the MC simulations repeated with $n = 5000$. Similar conclusions are obtained, except the absolute errors are smaller with this larger sample size.

4 Application: inference in nonparametric demand estimation with endogeneity

We now turn to inference on policy-relevant welfare functionals in nonparametric demand estimation. Following a large literature on nonparametric demand estimation (see, e.g., Hausman and Newey (1995); Vanhems (2010); Blundell et al. (2012); Blundell, Horowitz, and Parey (2013) and references therein), we assume that the demand of consumer i for some good is given by:

$$\mathsf{Q}_i \quad = \quad h_0(\mathsf{P}_i, \mathsf{Y}_i) + \mathsf{u}_i \tag{9}$$

where Q_i is the quantity of some good demanded, P_i is the price paid, and Y_i is the income of consumer i, and u_i is an error term.[12] Hausman and Newey (1995) provided limit theory for consumer surplus and deadweight loss functionals of the nonparametric demand function h_0 assuming prices and incomes are *exogenous*. In certain settings it is reasonable to allow prices, and possibly incomes, to be endogenous. One example is estimation the of gasoline demand from household-level data (Schmalensee and Stoker, 1999; Yatchew and No, 2001; Blundell et al., 2012, 2013). Even with household-level data there is evidence of endogeneity in prices (Yatchew and No, 2001; Blundell et al., 2013). Gasoline prices in a small local area and distance to the Gulf Coast have been suggested as instruments for gasoline price (see Yatchew and No (2001) and Blundell et al. (2013), respectively). In this case, model (9) falls into the class of models discussed in Section 2.5. Another example is the estimation of static models of labor supply, in which Q_i represents hours worked, P_i is the wage, and Y_i is other income. In this setting it is reasonable to allow for endogeneity of both

[12] We have followed Blundell et al. (2012) in modeling Q as the dependent variable, but the following analysis can easily be extended to take some transformation of Q as the dependent variable, as in Hausman and Newey (1995).

P_i and Y_i (see Blundell, Duncan, and Meghir (1998), Blundell, MaCurdy, and Meghir (2007), and references therein). Therefore, we extend the analysis of Hausman and Newey (1995) to allow for potential endogeneity in prices and incomes.

Previously, Vanhems (2010) established convergence rates for plug-in estimators of consumer surplus allowing for endogeneity of prices. More recently, Blundell et al. (2012) nonparametrically estimated exact deadweight loss from a first-stage kernel-based estimate of the demand function h_0 (subject to the Slutsky inequality restriction) allowing for endogeneity of prices. Neither of these papers established asymptotic distribution of their estimators.

The first functional of interest is exact consumer surplus (CS), namely the equivalent variation of a price change from p^0 to p^1 at income level y (fixed over the price change), which we denote by $\mathsf{S}_\mathsf{y}(\mathsf{p}^0)$. Let $\mathsf{p} : [0,1] \to \mathbb{R}$ denote a twice continuously differentiable path with $\mathsf{p}(0) = \mathsf{p}^0$ and $\mathsf{p}(1) = \mathsf{p}^1$. Hausman (1981) shows that $\mathsf{S}_\mathsf{y}(\mathsf{p}^0)$ is the solution to

$$\begin{aligned} \frac{\partial \mathsf{S}_\mathsf{y}(\mathsf{p}(t))}{\partial t} &= -h_0\big(\mathsf{p}(t), \mathsf{y} - \mathsf{S}_\mathsf{y}(\mathsf{p}(t))\big)\frac{\mathrm{d}\mathsf{p}(t)}{\mathrm{d}t} \\ \mathsf{S}_\mathsf{y}(\mathsf{p}(1)) &= 0\,. \end{aligned} \tag{10}$$

The second functional of interest is the deadweight loss (DWL) of the price change from p^0 to p^1 at income level y:

$$\mathsf{D}_\mathsf{y}(\mathsf{p}^0) = \mathsf{S}_\mathsf{y}(\mathsf{p}^0) - (\mathsf{p}^1 - \mathsf{p}^0)h_0(\mathsf{p}^1, \mathsf{y})\,. \tag{11}$$

In what follows we use the notation

$$\begin{aligned} f_{CS}(h) &= \text{solution to (10) with } h \text{ in place of } h_0 \\ f_{DWL}(h) &= f_{CS}(h) - (\mathsf{p}^1 - \mathsf{p}^0)h(\mathsf{p}^1, \mathsf{y})\,. \end{aligned}$$

In this notation we have $\mathsf{S}_\mathsf{y}(\mathsf{p}^0) = f_{CS}(h_0)$ and $\mathsf{D}_\mathsf{y}(\mathsf{p}^0) = f_{DWL}(h_0)$. We estimate CS and DWL using the plug-in estimators $f_{CS}(\widehat{h})$ and $f_{DWL}(\widehat{h})$.[13]

As is evident from Hausman and Newey (1995), sup-norm convergence rates of $\widehat{h}$ and its derivatives are required to control the nonlinearity bias when estimating CS and DWL using the plug-in estimators $f_{CS}(\widehat{h})$ and $f_{DWL}(\widehat{h})$.[14]

Both CS and DWL will typically be irregular (i.e. slower than $\sqrt{n}$-estimable) functionals of h_0 when prices and incomes are allowed to be endogenous. Inference on CS and DWL may be performed

[13] Modulo other considerations, the functional form of our DWL estimator is different from that used recently by Blundell et al. (2012), namely $\widehat{e}(\mathsf{p}^1) - \widehat{e}(\mathsf{p}^0) - (\mathsf{p}^1 - \mathsf{p}^0)\widehat{h}(\mathsf{p}^1, \widehat{e}(\mathsf{p}^1))$ where $\widehat{e}(\mathsf{p})$ is an estimated expenditure function which is obtained as the solution to a differential equation which, up to a change of sign, is the same as (10).

[14] The exception is when demand is independent of income, i.e. $h_0(\mathsf{p}, \mathsf{y}) = h_0(\mathsf{p})$, in which case $S_\mathsf{y}(\mathsf{p}^0)$ and $D_\mathsf{y}(\mathsf{p}^0)$ are linear functionals of h_0.

using studentized sieve t-statistics. To estimate the sieve variance of $f_{CS}(\widehat{h})$ and $f_{DWL}(\widehat{h})$, define

$$\begin{aligned}\frac{\partial f_{CS}(\widehat{h})}{\partial h}[\psi^J] &= \int_0^1 \left(\psi^J(\mathsf{p}(t),\mathsf{y}-\widehat{\mathsf{S}}_{\mathsf{y}}(t))e^{-\int_0^t \partial_2\widehat{h}(\mathsf{p}(v),\mathsf{y}-\widehat{\mathsf{S}}_{\mathsf{y}}(v))\mathsf{p}'(v)\,\mathrm{d}v}\mathsf{p}'(t)\right)\mathrm{d}t \\ \frac{\partial f_{DWL}(\widehat{h})}{\partial h}[\psi^J] &= \frac{\partial f_{CS}(\widehat{h})}{\partial h}[\psi^J] + (\mathsf{p}^1-\mathsf{p}^0)\psi^J(\mathsf{p}^1,\mathsf{y})\end{aligned}$$

where $\mathsf{p}'(t) = \frac{\mathrm{d}\mathsf{p}(t)}{\mathrm{d}t}$ and $\partial_2 h$ denotes the partial derivative of h with respect to its second argument and $\widehat{\mathsf{S}}_y(t)$ denotes the solution to (10) with $\widehat{h}$ in place of h_0. The sieve variances of f_{CS} and f_{DWL} are

$$\begin{aligned}\widehat{V}_{CS,n} &= \frac{\partial f_{CS}(\widehat{h})}{\partial h}[\psi^J]'[\widehat{S}'\widehat{G}_b^{-1}\widehat{S}]^{-1}\widehat{S}'\widehat{G}_b^{-1}\widehat{\Omega}\widehat{G}_b^{-1}\widehat{S}[\widehat{S}'\widehat{G}_b^{-1}\widehat{S}]^{-1}\frac{\partial f_{CS}(\widehat{h})}{\partial h}[\psi^J] \\ \widehat{V}_{DWL,n} &= \frac{\partial f_{DWL}(\widehat{h})}{\partial h}[\psi^J]'[\widehat{S}'\widehat{G}_b^{-1}\widehat{S}]^{-1}\widehat{S}'\widehat{G}_b^{-1}\widehat{\Omega}\widehat{G}_b^{-1}\widehat{S}[\widehat{S}'\widehat{G}_b^{-1}\widehat{S}]^{-1}\frac{\partial f_{DWL}(\widehat{h})}{\partial h}[\psi^J]\end{aligned}$$

where $\widehat{S} = B'\Psi/n$, $\widehat{G}_b = B'B/n$, and $\widehat{\Omega} = n^{-1}\sum_{i=1}^n \widehat{\mathsf{u}}_i^2 b^K(\mathsf{W}_i)b^K(\mathsf{W}_i)'$ with $\widehat{\mathsf{u}}_i = (\mathsf{Q}_i - \widehat{h}(\mathsf{X}_i))$ and $\mathsf{X}_i = (\mathsf{P}_i,\mathsf{Y}_i)'$. We take W_i to be a 2×1 vector of instruments when P_i and Y_i are endogenous, and we take $\mathsf{W}_i = (\mathsf{W}_{1i},\mathsf{Y}_i)'$ when Y_i is exogenous where W_{1i} an instrument for P_i.

We now present regularity conditions under which sieve t-statistics for $f_{CS}(\widehat{h})$ and $f_{DWL}(\widehat{h})$ are asymptotically $N(0,1)$. In the case in which both P_i and Y_i are endogenous, let $T: L^2(\mathsf{X}) \to L^2(\mathsf{W})$ be compact with singular value decomposition $\{\phi_{0j},\phi_{1j},\mu_j\}_{j=1}^\infty$ where

$$T\phi_{0j} = \mu_j\phi_{1j}, \quad (T^*T)^{1/2}\phi_{0j} = \mu_j\phi_{0j}, \quad (TT^*)^{1/2}\phi_{1j} = \mu_j\phi_{1j}$$

and $\{\phi_{0j}\}_{j=1}^\infty$ and $\{\phi_{0j}\}_{j=1}^\infty$ are orthonormal bases for $L^2(\mathsf{X})$ and $L^2(\mathsf{W})$, respectively. In the case in which P_i is endogenous but Y_i is exogenous, we let $T_{\mathsf{y}}: L^2(\mathsf{P}|\mathsf{Y}=\mathsf{y}) \to L^2(\mathsf{W}_1|\mathsf{Y}=\mathsf{y})$ be compact with singular value decomposition $\{\phi_{0j,\mathsf{y}},\phi_{1j,\mathsf{y}},\mu_{j,\mathsf{y}}\}_{j=1}^\infty$ for each y where

$$T_{\mathsf{y}}\phi_{0j,\mathsf{y}} = \mu_{j,\mathsf{y}}\phi_{1j,\mathsf{y}}, \quad (T_{\mathsf{y}}^*T_{\mathsf{y}})^{1/2}\phi_{0j,\mathsf{y}} = \mu_{j,\mathsf{y}}\phi_{0j,\mathsf{y}}, \quad (T_{\mathsf{y}}T_{\mathsf{y}}^*)^{1/2}\phi_{1j,\mathsf{y}} = \mu_{j,\mathsf{y}}\phi_{1j,\mathsf{y}}$$

and $\{\phi_{0j,\mathsf{y}}\}_{j=1}^\infty$ and $\{\phi_{0j,\mathsf{y}}\}_{j=1}^\infty$ are orthonormal bases for $L^2(\mathsf{P}|\mathsf{Y}=\mathsf{y})$ and $L^2(\mathsf{W}_1|\mathsf{Y}=\mathsf{y})$, respectively. In this case, we define $\phi_{0j}(\mathsf{p},\mathsf{y}) = \phi_{0j,\mathsf{y}}(\mathsf{p})$, $\phi_{1j}(\mathsf{w}_1,y) = \phi_{1j,\mathsf{y}}(\mathsf{w}_1)$, and $\mu_j^2 = E[\mu_{j,\mathsf{Y}_i}^2]$ (see Section 2.5 for further details). In both cases, for fixed $\mathsf{p}^0,\mathsf{p}^1,\mathsf{y}$ we define

$$a_j = a_j(\mathsf{p}^0,\mathsf{p}^1,\mathsf{y}) = \int_0^1 \left(\phi_{0j}(\mathsf{p}(t),\mathsf{y}-S_{\mathsf{y}}(t))e^{-\int_0^t \partial_2 h_0(\mathsf{p}(v),\mathsf{y}-S_{\mathsf{y}}(v))\mathsf{p}'(v)\,\mathrm{d}v}\mathsf{p}'(t)\right)\mathrm{d}t \tag{12}$$

for each $j \geq 1$. We also follow Chen and Pouzo (2014) and assume that Ψ_J and B_K are Riesz bases in that they span $\phi_{01},\ldots,\phi_{0J}$ and $\phi_{11},\ldots,\phi_{1J}$, respectively.

Assumption CS *(i)* X_i *and* W_i *both have compact rectangular support and densities bounded away from* 0 *and* ∞*; (ii)* $h_0 \in B_\infty(p, L)$ *with* $p > 2$ *and* $0 < L < \infty$*; (iii)* $E[\mathsf{u}_i^2|\mathsf{W}_i = w]$ *is uniformly bounded away from* 0 *and* ∞*,* $E[|\mathsf{u}_i|^{2+\delta}]$ *is finite, and* $\sup_w E[u_i^2\{|u_i| > \ell(n)\}|W_i = w] = o(1)$ *for any positive sequence with* $\ell(n) \nearrow \infty$*; (iv)* Ψ_J *is spanned by a (tensor-product) B-spline basis of order* $\gamma > p$ *or continuously differentiable wavelet basis of regularity* $\gamma > p$ *and* B_K *is spanned by a (tensor-product) B-spline, wavelet or cosine basis; and (v)*

$$\frac{\sqrt{n}}{\left(\sum_{j=1}^J (a_j/\mu_j)^2\right)^{1/2}} \times \left(J^{-p/2} + \left(J^{-p/2} + \mu_J^{-1}\sqrt{(J\log n)/n}\right)^2 (1 + J^{1/2})\right) = o(1)\,,$$

$(J^{(2+\delta)/(2\delta)} \vee J^{3/2}\mu_J^{-1} \vee J^2\mu_J^{-2}(\sum_{j=1}^J (a_j/\mu_j)^2)^{-1/2})\sqrt{(\log n)/n} = o(1)$.

Assumption CS(i)–(iv) is very similar to Assumptions 1, 2, 7, and 8 in Hausman and Newey (1995) with the exception that Hausman and Newey (1995) use a power series basis and work with log prices, log incomes, and log demand.

Our first result is asymptotic normality of sieve t-statistics for CS functionals.

Theorem 4.1 *Let Assumption CS hold. Then:*

$$\sqrt{n}\frac{f_{CS}(\widehat{h}) - f_{CS}(h_0)}{\widehat{V}_{CS,n}^{1/2}} \Rightarrow N(0,1)\,.$$

We now present a corresponding result for DWL functionals. To introduce the result, define $\bar{a}_j = a_j + (\mathsf{p}^1 - \mathsf{p}^0)\phi_{0j}(\mathsf{p}^1, \mathsf{y})$.

Theorem 4.2 *Let Assumption CS hold with* $\bar{a}_j$ *in place of* a_j*. Then:*

$$\sqrt{n}\frac{f_{DWL}(\widehat{h}) - f_{DWL}(h_0)}{\widehat{V}_{DWL,n}^{1/2}} \Rightarrow N(0,1)\,.$$

Hausman and Newey (1995) suggest that $a_j = o(\bar{a}_j)$ as $j \to \infty$ because a_j is a smooth integral functional whereas $\bar{a}_j$ depends on the functions evaluated at a particular point. Therefore we should expect that the convergence rate of $f_{DWL}(\widehat{h})$ to be slower than that of $f_{CS}(\widehat{h})$. For this reason we do not derive the joint asymptotic distribution of $f_{CS}(\widehat{h})$ and $f_{DWL}(\widehat{h})$.

We now provide more concrete conditions under which the sieve t-statistics for exact consumer surplus are asymptotically normal, allowing for endogeneity. Analogous results hold for deadweight loss if we replace a_j by $\bar{a}_j$ in what follows.

Corollary 4.1 *Let Assumption CS(i)–(iv) hold and let $a_j \asymp j^{a/2}$. Then:*

(1) Mildly ill-posed case: let $\mu_j \asymp j^{-\varsigma/2}$ with $a + \varsigma \geq -1$ and $\delta \geq 2/(2 + \varsigma - (a \wedge 0))$ and let $nJ^{-(p+a+\varsigma+1)} = o(1)$ and $J^{3+\varsigma-(a\wedge 0)}(\log n)/n = o(1)$. Then: the population counterpart $V_{CS,n}$ of $\widehat{V}_{CS,n}$ behaves like

$$V_{CS,n} \asymp \textstyle\sum_{j=1}^{J}(a_j/\mu_j)^2 \asymp \sum_{j=1}^{J} j^{a+\varsigma} \asymp J^{(a+\varsigma)+1}\,,$$

Assumption CS(v) holds, and the sieve t-statistic for $f_{CS}(h_0)$ is asymptotically $N(0,1)$.

(2) Severely ill-posed case: let $\mu_j \asymp \exp(-\frac{1}{2}j^{\varsigma/2})$ and $a > 0$ and take $J = (\log(n/(\log n)^{\varrho}))^{2/\varsigma}$ where $\varrho > 0$ is chosen such that $\varrho\varsigma > (6+\varsigma) \vee (8-2a+\varsigma) \vee (6+2\varsigma-2a)$. Then: the population counterpart $V_{CS,n}$ of $\widehat{V}_{CS,n}$ behaves like

$$V_{CS,n} \gtrsim \frac{n}{(\log n)^{\varrho}} \times (\log(n/(\log n)^{\varrho}))^{2a/\varsigma}\,,$$

Assumption CS(v) holds, and the sieve t-statistic for $f_{CS}(h_0)$ is asymptotically $N(0,1)$.

Note that we may choose J satisfying the stated conditions for the mildly ill-posed case provided $p > 2 - a - (a \wedge 0)$, which is trivially true if $p > 2$ whenever $a \geq 0$. We may also choose such a ϱ for the severely ill-posed case whenever $4p > [(6 + \varsigma) \vee (8 - 2a + \varsigma) \vee (6 + 2\varsigma - 2a)] - 2a + 2$.

We finish this section by stating conditions for asymptotic normality of $f_{CS}(\widehat{h})$ in the exogenous case in which $\tau_J = 1$ and the sieve NPIV estimator reduces to the usual series LS estimator. Let the basis functions span an orthonormal basis $\phi_1, \ldots, \phi_J$ for each J and let a_j be as defined in (12) with ϕ_j in place of ϕ_{0j}. Assumption CS(v) then applies with $\mu_j = 1$ for each j. The following result describes the regularity conditions for asymptotic normality of $f_{CS}(\widehat{h})$. Analogous results for $f_{DWL}(\widehat{h})$ hold if we replace a_j by $\bar{a}_j = a_j + (\mathsf{p}^1 - \mathsf{p}^0)\phi_j(\mathsf{p}^1, \mathsf{y})$ in what follows.

Corollary 4.2 *Let Assumption CS(i)–(iv) hold, let $a_j \asymp j^{a/2}$ with $a \geq -1$, and let $nJ^{-(p+a+1)} = o(1)$ and $J^{3-(a\wedge 0)}(\log n)/n = o(1)$ and $\delta \geq 2/(2 - (a \wedge 0))$. Then: the population counterpart $V_{CS,n}$ of $\widehat{V}_{CS,n}$ behaves like $V_{CS,n} \asymp J^{a+1}$, Assumption CS(v) holds, and the sieve t-statistic for $f_{CS}(h_0)$ is asymptotically $N(0,1)$.*

Hausman and Newey (1995) establish asymptotic normality of t-statistics for exact CS and DWL plug-in estimators based on a kernel estimator of demand. They also establish asymptotic normality of t-statistics for *averaged* exact CS and DWL plug-in estimators (i.e. exact CS/DWL averaged over a range of incomes) based on a series LS estimator of demand with power series basis, assuming h_0 to is infinitely times differentiable and $J^{22}/n = o(1)$. Newey (1997) establishes asymptotic normality of t-statistics for *approximate* CS functionals based on series LS estimators of demand under weaker conditions (i.e. $nJ^{-p} = o(1)$ and either $J^6/n = o(1)$ for power series or $J^4/n = o(1)$ for splines), but

also without endogeneity.[15] Corollary 4.2 complements their analysis by providing conditions for asymptotic normality of exact CS and DWL functionals based on series LS estimators of demand.

Appendix A Pointwise asymptotic normality of sieve t-statistics

In this section we establish asymptotic normality of sieve t-statistics for $f(h_0)$ where $f : L^\infty(X) \to \mathbb{R}$ is any nonlinear functional of NPIV. Under some high-level conditions, Chen and Pouzo (2014) established the pointwise asymptotic normality of the sieve t statistics for (possibly) nonlinear functionals of h_0 satisfying general semi/nonparametric conditional moment restrictions including NPIV and nonparametric quantile IV as special cases. As the sieve NPIV estimator $\widehat{h}$ has a closed-form expression (unlike, say, nonparametric quantile IV) we derive the limit theory directly rather than appealing to the general theory in Chen and Pouzo (2014). Our regularity conditions are tailored to the case in which $f(h_0)$ is *irregular* (i.e. slower than root-n estimable).

Denote the derivative of f at h_0 in the direction $v \in \mathcal{V} := (L^2(X) - \{h_0\})$ by

$$\frac{\partial f(h_0)}{\partial h}[g] = \lim_{\delta \to 0^+} \frac{f(h_0 + \delta g)}{\delta} .$$

If f is a linear functional then $\frac{\partial f(h_0)}{\partial h}[g] = f(g)$. The sieve 2SLS Riesz representer of $\frac{\partial f(h_0)}{\partial h}$ is

$$v_n^*(x) = \psi^J(x)'[S'G_b^{-1}S]^{-1}\frac{\partial f(h_0)}{\partial h}[\psi^J]$$

where $\frac{\partial f(h_0)}{\partial h}[\psi^J]$ denotes the vector formed by applying $\frac{\partial f(h_0)}{\partial h}[\cdot]$ to each element of ψ^J. Define the weak norm $\|\cdot\|$ on Ψ_J as $\|h\| = \|\Pi_K T h\|_{L^2(W)}$. Then

$$\|v_n^*\|^2 = \frac{\partial f(h_0)}{\partial h}[\psi^J]'[S'G_b^{-1}S]^{-1}\frac{\partial f(h_0)}{\partial h}[\psi^J] .$$

We say that the functional f is an irregular (i.e. slower than $\sqrt{n}$-estimable) functional of h_0 if $\|v_n^*\| \nearrow \infty$ and a regular (i.e. $\sqrt{n}$-estimable) functional of h_0 if $\|v_n^*\| \nearrow \|v^*\| < \infty$.

The sieve 2SLS variance $\|v_n^*\|_{sd}^2$ is defined as

$$\|v_n^*\|_{sd}^2 = \frac{\partial f(h_0)}{\partial h}[\psi^J]'[S'G_b^{-1}S]^{-1}S'G_b^{-1}\Omega G_b^{-1}S[S'G_b^{-1}S]^{-1}\frac{\partial f(h_0)}{\partial h}[\psi^J]$$

[15] Using the results in Chen and Christensen (2014), one can show that sieve t-statistics for *approximate* CS based on spline or wavelet LS estimates of log demand without endogeneity are asymptotically normal under assumptions comparable to Assumption CS(i)–(iii) provided $nJ^{-(p+\alpha+1)} = o(1)$, $J^{1-(a\wedge 0)}(\log n)^2/n = o(1)$, and $J^{(2+\delta)/\delta}(\log n)/n = o(1)$.

where $\Omega = \Omega_K = E[u_i^2 b^K(W_i) b^K(W_i)']$. Our estimator of $\|v_n^*\|_{sd}^2$ is

$$\widehat{\|v_n^*\|}_{sd}^2 = \frac{\partial f(\widehat{h})}{\partial h}[\psi^J]'[\widehat{S}'\widehat{G}_b^{-1}\widehat{S}]^{-1}\widehat{S}'\widehat{G}_b^{-1}\widehat{\Omega}\widehat{G}_b^{-1}\widehat{S}[\widehat{S}'\widehat{G}_b^{-1}\widehat{S}]^{-1}\frac{\partial f(\widehat{h})}{\partial h}[\psi^J]$$

where $\widehat{\Omega} = n^{-1}\sum_{i=1}^n \widehat{u}_i^2 b^K(W_i) b^K(W_i)'$ with $\widehat{u}_i = (Y_i - \widehat{h}(X_i))$, which is analogous to the usual linear 2SLS variance estimator. The scaled sieve Riesz representer

$$u_n^* = v_n^*/\|v_n^*\|_{sd}$$

has the property that $\|u_n^*\| \asymp 1$ irrespective of whether $f(h_0)$ is regular or irregular. Finally, denote

$$\widehat{v}_n^*(x) = \psi^J(x)'[S'G_b^{-1}S]^{-1}\frac{\partial f(\widehat{h})}{\partial h}[\psi^J]$$

where clearly $v_n^* = \widehat{v}_n^*$ whenever $f(\cdot)$ is linear.

Assumption 2 (continued) *(iii)* $E[u_i^2|W_i = w] \geq \underline{\sigma}^2 > 0$ *uniformly for* $w \in \mathcal{W}$*; and (iv)* $\sup_w E[u_i^2\{|u_i| > \ell(n)\}|W_i = w] = o(1)$ *for any positive sequence with* $\ell(n) \nearrow \infty$.

Assumption 5 *Either (a) or (b) of the following hold:*

(a) f *is a linear functional and* $\|v_n^*\|^{-1}(f(\widetilde{h}) - f(h_0)) = o_p(n^{-1/2})$*; or*

(b) (i) $g \mapsto \frac{\partial f(h_0)}{\partial h}[g]$ *is a linear functional; (ii)*

$$\sqrt{n}\frac{(f(\widehat{h}) - f(h_0)}{\|v_n^*\|} = \sqrt{n}\frac{\frac{\partial f(h_0)}{\partial h}[\widehat{h} - \widetilde{h}]}{\|v_n^*\|} + o_p(1)\,;$$

and (iii) $\frac{\|\widehat{v}_n^* - v_n^*\|}{\|v_n^*\|} = o_p(1)$.

Assumption 2(iv) is a mild condition which is trivially satisfied if $E[|u_i|^{2+\epsilon}|W_i = w]$ is uniformly bounded for some $\epsilon > 0$. Assumption 5(a) and (b)(i)(ii) is similar to Assumption 3.5 of Chen and Pouzo (2014). Assumption 5(b)(iii) controls any additional error arising in the estimation of $\widehat{\|v_n^*\|}_{sd}$ due to nonlinearity of $f(\cdot)$ and is not required when $f(\cdot)$ is a linear functional. Previously, Chen and Pouzo (2014) verified their Assumption 3.5 using a plug-in sieve minimum distance estimator of a weighted quadratic functional example. However, without a sup-norm convergence rate, it could be difficult to verify the high-level conditions for nonlinear functionals that are more complicated than a quadratic functional of NPIV.

Remark A.1 *Let* $\mathcal{H}_n \subseteq \mathcal{H}$ *be a neighborhood of* h_0 *such that* $\widehat{h}, \widetilde{h} \in \mathcal{H}_n$ *wpa1. Sufficient conditions for Assumptions 5(a) and (b)(i)(ii) are:*

(a') *(i) f is a linear functional and there exists α with $|\alpha| \geq 0$ s.t. $|f(h-h_0)| \lesssim \|\partial^\alpha h - \partial^\alpha h_0\|_\infty$ for all $h \in \mathcal{H}_n$; and (ii) $\|v_n^*\|^{-1}\|\partial^\alpha \widetilde{h} - \partial^\alpha h_0\|_\infty = o_p(n^{-1/2})$; or*

(b') *(i) $g \mapsto \frac{\partial f(h_0)}{\partial h}[g]$ is a linear functional and there exists α with $|\alpha| \geq 0$ s.t. $|\frac{\partial f(h_0)}{\partial h}[h-h_0]| \lesssim \|\partial^\alpha h - \partial^\alpha h_0\|_\infty$ for all $h \in \mathcal{H}_n$; (ii) there exists α_1, α_2 with $|\alpha_1|, |\alpha_2| \geq 0$ s.t.*

$$\left| f(\widehat{h}) - f(h_0) - \frac{\partial f(h_0)}{\partial h}[\widehat{h} - h_0] \right| \lesssim \|\partial^{\alpha_1}\widehat{h} - \partial^{\alpha_1} h_0\|_\infty \|\partial^{\alpha_2}\widehat{h} - \partial^{\alpha_2} h_0\|_\infty ;$$

and (iii) $\|v_n^*\|^{-1}(\|\partial^{\alpha_1}\widehat{h} - \partial^{\alpha_1} h_0\|_\infty \|\partial^{\alpha_2}\widehat{h} - \partial^{\alpha_2} h_0\|_\infty + \|\partial^\alpha \widetilde{h} - \partial^\alpha h_0\|_\infty) = o_p(n^{-1/2})$.

Condition (a')(i) is trivially satisfied for any evaluation functional of the form $f(h) = \partial^\alpha h(\bar{x})$ for fixed $\bar{x} \in \mathcal{X}$ with $\mathcal{H}_n = \mathcal{H}$. Condition (b')(i)(ii) are satisfied by typical nonlinear welfare functionals such as exact consumer surplus and deadweight loss functionals (see Hausman and Newey (1995)). Conditions (a')(i) and (b')(iii) can be verified given the sup-norm rate results in Section 2.

Theorem A.1 *(1) Let Assumptions 1(iii), 2(i)(iii)(iv), 4(i), and either 5(a) or 5(b)(i)(ii) hold, and let $\tau_J \zeta \sqrt{(J \log n)/n} = o(1)$. Then:*

$$\sqrt{n}\frac{(f(\widehat{h}) - f(h_0))}{\|v_n^*\|_{sd}} \Rightarrow N(0,1) .$$

(2) If $\|\widehat{h} - h_0\|_\infty = o_p(1)$ and Assumptions 2(ii) and 3(iii) hold (and 5(b)(iii) also holds if f is nonlinear), then:

$$\begin{aligned} \left| \frac{\widehat{\|v_n^*\|}_{sd}}{\|v_n^*\|_{sd}} - 1 \right| &= o_p(1) \\ \sqrt{n}\frac{(f(\widehat{h}) - f(h_0))}{\widehat{\|v_n^*\|}_{sd}} &\Rightarrow N(0,1) . \end{aligned}$$

Chen and Pouzo (2014) establish asymptotic normality of plug-in estimators of possibly nonlinear functionals in general semi/nonparametric conditional moment restriction models with endogeneity. By exploiting the close form expression of the sieve NPIV estimator and by applying exponential inequalities for random matrices, Theorem A.1 derives the limit theory allowing for faster growth rate of J than Remark 6.1 in Chen and Pouzo (2014).

Appendix B Bootstrap uniform limit theory for sieve t-statistics

We now show how our sup-norm rate results and tight bounds on random matrices can be used to derive bootstrap uniform confidence bands for a class of general nonlinear functionals $\{f_t(\cdot) : t \in \mathcal{T}\}$

of h_0 in a NPIV model where $\mathcal{T}$ is a (possibly infinite dimensional) index set. In particular, we establish validity of a sieve score bootstrap for estimating the distribution of supremum of the sieve t-statistic process (i.e. the process formed by calculating the sieve t-statistic for each $f_t(h_0)$), which leads to asymptotically exact uniform confidence bands for $\{f_t(h_0) : t \in \mathcal{T}\}$.

Let $\mathcal{T}$ be a closed subset of a separable metric space and let $f_t : L^\infty(X) \to \mathbb{R}$ for each $t \in \mathcal{T}$. For each $t \in \mathcal{T}$ we define

$$\begin{aligned}
\frac{\partial f_t(h_0)}{\partial h}[g] &= \lim_{\delta \to 0^+} \frac{f_t(h_0 + \delta g)}{\delta} \\
v^*_{n,t}(x) &= \psi^J(x)'[S'G_b^{-1}S]^{-1}\frac{\partial f_t(h_0)}{\partial h}[\psi^J] \\
\widehat{v}^*_{n,t}(x) &= \psi^J(x)'[S'G_b^{-1}S]^{-1}\frac{\partial f_t(\widehat{h})}{\partial h}[\psi^J] \\
\|v^*_{n,t}\|^2 &= \frac{\partial f_t(h_0)}{\partial h}[\psi^J]'[S'G_b^{-1}S]^{-1}\frac{\partial f_t(h_0)}{\partial h}[\psi^J] \\
\|v^*_{n,t}\|^2_{sd} &= \frac{\partial f_t(h_0)}{\partial h}[\psi^J]'[S'G_b^{-1}S]^{-1}S'G_b^{-1}\Omega G_b^{-1}S[S'G_b^{-1}S]^{-1}\frac{\partial f_t(h_0)}{\partial h}[\psi^J] \\
\widehat{\|v^*_{n,t}\|}^2_{sd} &= \frac{\partial f_t(\widehat{h})}{\partial h}[\psi^J]'[\widehat{S}'\widehat{G}_b^{-1}\widehat{S}]^{-1}\widehat{S}'\widehat{G}_b^{-1}\widehat{\Omega}\widehat{G}_b^{-1}\widehat{S}[\widehat{S}'\widehat{G}_b^{-1}\widehat{S}]^{-1}\frac{\partial f_t(\widehat{h})}{\partial h}[\psi^J] \\
u^*_{n,t}(x) &= v^*_{n,t}(x)/\|v^*_{n,t}\|_{sd}
\end{aligned}$$

with $\|\cdot\|$, Ω, and $\widehat{\Omega}$ as defined in Appendix A.

To construct uniform confidence bands for $\{f_t(h_0) : t \in \mathcal{T}\}$ we propose the following sieve score bootstrap procedure. Let $\varpi_1, \ldots, \varpi_n$ be a bootstrap sample of IID random variables drawn independently of the data, with $E[\varpi_i|Z^n] = 0$, $E[\varpi_i^2|Z^n] = 1$, $E[|\varpi_i|^{2+\epsilon}|Z^n] < \infty$ for some $\epsilon \geq 1$. Common examples of distributions for ϖ_i include $N(0,1)$, recentered exponential, Rademacher, or the two-point distribution of Mammen (1993).[16] The sieve score bootstrap process $\{\mathbb{Z}^*_n(t) :\in \mathcal{T}\}$ is given by

$$\mathbb{Z}^*_n(t) = \frac{\frac{\partial f_t(\widehat{h})}{\partial h}[\psi^J]'[\widehat{S}'\widehat{G}_b^{-1}\widehat{S}]^{-1}\widehat{S}'\widehat{G}_b^{-1}}{\widehat{\|v^*_{n,t}\|}_{sd}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n b^K(W_i)\widehat{u}_i\varpi_i\right)$$

for each $t \in \mathcal{T}$, where $\widehat{u}_i = Y_i - \widehat{h}(X_i)$.

Assumption 2 *(iv')* $\sup_w E[|u_i|^3|W_i = w] < \infty$.

Assumption 5' *Let η_n and η'_n be sequences of positive constants such that $\eta_n = o(1)$ and $\eta'_n = o(1)$. Either (a) or (b) of the following holds:*

[16] To generate recentered exponential weights let e_i be a draw from the exponential distribution with mean 1 and let $\varpi_i = e_i - 1$. The Rademacher weights put probability 0.5 on both 1 and -1. Mammen's two-point distribution puts probability $\frac{\sqrt{5}+1}{2\sqrt{5}}$ on $\frac{1-\sqrt{5}}{2}$ and probability $1 - \frac{\sqrt{5}+1}{2\sqrt{5}}$ on $\frac{\sqrt{5}+1}{\sqrt{2}}$.

(a) f_t *is a linear functional for each* $t \in \mathcal{T}$ *and* $\sup_{t\in\mathcal{T}} \sqrt{n}\|v^*_{t,n}\|^{-1}|f_t(\widetilde{h}) - f_t(h_0)| = O_p(\eta_n)$*; or*

(b) (i) $g \mapsto \frac{\partial f_t(h_0)}{\partial h}[g]$ *is a linear functional with* $\|v^*_{n,t}\| \neq 0$ *for each* $t \in \mathcal{T}$*; (ii)*

$$\sup_{t\in\mathcal{T}} \left| \sqrt{n}\frac{(f_t(\widehat{h}) - f(h_0)}{\|v^*_{n,t}\|} - \sqrt{n}\frac{\frac{\partial f_t(h_0)}{\partial h}[\widehat{h} - \widetilde{h}]}{\|v^*_{n,t}\|} \right| = O_p(\eta_n)\,;$$

and (iii) $\sup_{t\in\mathcal{T}} \frac{\|\widehat{v}^*_{n,t} - v^*_{n,t}\|}{\|v^*_{n,t}\|} = O_p(\eta'_n)$.

Let d_n denote the intrinsic semi-metric on $\mathcal{T}$ given by $d_n(t_1,t_2)^2 = E[(u^*_{n,t_1}(X_i) - u^*_{n,t_2}(X_i))^2]$ and let $N(\mathcal{T}, d_n, \epsilon)$ denote the ϵ-entropy of $\mathcal{T}$ with respect to d_n. Let $\delta_{h,n}$ be a sequence of positive constants such that $\delta_{h,n} = o(1)$ and define $\delta_{V,n} = (\zeta_{b,K}^{(2+\delta)/\delta}\sqrt{(\log K)/n})^{\delta/(1+\delta)} + \tau_J\zeta\sqrt{(\log J)/n})$.

Assumption 6 *(i)* $(\mathcal{T}, d_n)$ *is separable for each* n*;*
(ii) there exists a sequence of finite positive constants c_n *such that*

$$1 + \int_0^\infty \sqrt{\log N(\mathcal{T}, d_n, \epsilon)}\,\mathrm{d}\epsilon = O(c_n)\,;$$

and (iii) there exists a sequence of positive constants r_n *with* $r_n = o(1)$ *such that*

$$\frac{\zeta_{b,K}J^2}{r_n^3\sqrt{n}} = o(1)$$

and

$$\eta_n + \eta'_n\sqrt{J} + r_n + (\delta_{V,n} + \delta_{h,n} + \eta'_n) \times c_n = o(c_n^{-1})$$

where $\|\widehat{h} - h_0\|_\infty = O_p(\delta_{h,n}) = o_p(1)$ *and* $\eta'_n \equiv 0$ *if the* f_t *are linear.*

Assumptions 2(iv') is a mild condition used to derive the uniform limit theory. Assumption 5' is a uniform (in t) version of Assumption 5. Assumption 5'(iii) is only required for consistent variance estimation. Assumption 6 is a mild regularity condition requiring the class $\{u^*_t : t \in \mathcal{T}\}$ not be too complex. This condition is used to place tight bounds on the supremum of the bootstrap t-statistic processes.

Remark B.1 *Let* $\mathcal{H}_n \subseteq \mathcal{H}$ *be a neighborhood of* h_0 *such that* $\widehat{h}, \widetilde{h} \in \mathcal{H}_n$ *wpa1 and let* $\underline{v}_n$ *be such that* $\inf_{t\in\mathcal{T}} \|v^*_{n,t}\|_{sd} \geq \underline{v}_n > 0$ *for each* n*. Sufficient conditions for Assumptions 5'(a) and (b)(i)(ii) are:*

(a') (i) f_t *is a linear functional for each* $t \in \mathcal{T}$ *and there exists* α *with* $|\alpha| \geq 0$ *s.t.* $\sup_t |f_t(h - h_0)| \lesssim \|\partial^\alpha h - \partial^\alpha h_0\|_\infty$ *for all* $h \in \mathcal{H}_n$*; and (ii)* $\underline{v}_n^{-1}\|\partial^\alpha\widetilde{h} - \partial^\alpha h_0\|_\infty = o_p(n^{-1/2})$*; or*

(b') *(i)* $g \mapsto \frac{\partial f_t(h_0)}{\partial h}[g]$ *is a linear functional for each* $t \in \mathcal{T}$ *and there exists* α *with* $|\alpha| \geq 0$ *s.t.* $\sup_t |\frac{\partial f_t(h_0)}{\partial h}[h-h_0]| \lesssim \|\partial^\alpha h - \partial^\alpha h_0\|_\infty$ *for all* $h \in \mathcal{H}_n$*; (ii) there exists* α_1, α_2 *with* $|\alpha_1|, |\alpha_2| \geq 0$ *s.t.*

$$\sup_t \left| f_t(\widehat{h}) - f_t(h_0) - \frac{\partial f_t(h_0)}{\partial h}[\widehat{h} - h_0] \right| \lesssim \|\partial^{\alpha_1}\widehat{h} - \partial^{\alpha_1} h_0\|_\infty \|\partial^{\alpha_2}\widehat{h} - \partial^{\alpha_2} h_0\|_\infty ;$$

and (iii) $\underline{v}_n^{-1}(\|\partial^{\alpha_1}\widehat{h} - \partial^{\alpha_1} h_0\|_\infty \|\partial^{\alpha_2}\widehat{h} - \partial^{\alpha_2} h_0\|_\infty + \|\partial^\alpha \widetilde{h} - \partial^\alpha h_0\|_\infty) = o_p(n^{-1/2})$.

Condition (a')(i) is satisfied for any evaluation functional of the form $f_t(h) = \partial^\alpha h(t)$ with $\mathcal{T} \subseteq \mathcal{X}$ and $\mathcal{H}_n = \mathcal{H}$.

Remark B.2 *Let* $\mathcal{T} \subset \mathbb{R}^{d_T}$ *be compact and let there exist sequence of positive constants* Γ_n, γ_n *such that*

$$\sup_{h \in \Psi_J : \|h\|_{L^2(X)} = 1} |f_{t_1}(h) - f_{t_2}(h)| \leq \Gamma_n \|t_1 - t_2\|_{\ell^2}^{\gamma_n}$$

if the f_t *are linear functionals, or*

$$\sup_{h \in \Psi_J : \|h\|_{L^2(X)} = 1} \left| \left(\frac{\partial f_{t_1}(h_0)}{\partial h}[h] - \frac{\partial f_{t_2}(h_0)}{\partial h}[h] \right) \right| \leq \Gamma_n \|t_1 - t_2\|_{\ell^2}^{\gamma_n}$$

if the f_t *are nonlinear, and let Assumption 1(iii) and 4(i) hold. Then: Assumption 6(i)(ii) holds with* $c_n = 1 + \int_0^\infty \sqrt{d_T \{\log(\tau_J \Gamma_n \epsilon^{-\gamma_n} / \underline{v}_n) \vee 0\}}\, \mathrm{d}\epsilon$.

Let $\mathbb{P}^*$ denote the probability measure of the bootstrap innovations $\varpi_1, \ldots, \varpi_n$ conditional on the data $Z^n := \{(X_1, Y_1, W_1), \ldots, (X_n, Y_n, W_n)\}$.

Theorem B.1 *Let Assumptions 1(iii), 2(i)–(iii)(iv'), 3(iii), 4(i), 5', and 6 hold and* $\tau_J \zeta \sqrt{(J \log n)/n} = o(1)$. *Then:*

$$\sup_{s \in \mathbb{R}} \left| \mathbb{P}\left(\sup_{t \in \mathcal{T}} \left| \frac{\sqrt{n}(f_t(\widehat{h}) - f_t(h_0))}{\widehat{\|v_{n,t}^*\|}_{sd}} \right| \leq s \right) - \mathbb{P}^*\left(\sup_{t \in \mathcal{T}} |\mathbb{Z}_n^*(t)| \leq s \right) \right| = o_p(1) .$$

Theorem B.1 establishes consistency of our sieve score bootstrap procedure for estimating the critical values of the uniform sieve t-statistic process for a NPIV model.

Remark B.3 *Theorem B.1 applies to uniform confidence bands for* $\partial^\alpha h_0$ *as a special case in which* $f_t(h) = h(t)$ *and* $\mathcal{T} \subseteq \mathcal{X}$ *provided Assumptions 1, 2(i)–(iii)(iv'), 3(iii), and 4 hold,* Ψ_J *is formed from a B-spline basis of regularity* $\gamma > (p \vee 2 + |\alpha|)$, B_K *is spanned by a B-spline, wavelet, or cosine basis,* $\|v_{n,t}^*\|_{sd} \asymp \tau_J J^a$ *for some* $a > 0$ *uniformly in* t, *and* $\tau_J J \sqrt{(\log n)/n} = o(1)$, $J^{-p/d} = o(\tau_J \sqrt{(J^{2a} \log n)/n})$, $J^5 (\log J)^6 / n = o(1)$, *and* $J^{(2+\delta)} (\log J)^{(1+2\delta)} = o(n^\delta)$.

Theorem B.1 contributes to the recent literature on inference for irregular (possibly) nonlinear functionals of nonparametric ill-posed inverse problems. For (pointwise) inference on irregular (possibly) nonlinear functionals of general semi/nonparametric conditional moment restrictions *with endogeneity*, Chen and Pouzo (2014) establish the validity of generalized residual bootstrap sieve t and sieve QLR statistics, and also present a sieve score bootstrap in their supplemental Appendix D. Horowitz and Lee (2012) were the first to derive uniform confidence bands for h_0 in a NPIV model based on the modified orthogonal series NPIV estimator of Horowitz (2011).[17] Our Theorem B.1 and Remark B.3 include uniform confidence bands for h_0 as a special case in which $f_t(h) = h(t)$ and $\mathcal{T} \subseteq \mathcal{X}$.[18] In an important paper on series least squares (LS) regression *without endogeneity*, Belloni et al. (2014) extend the Gaussian simulation (conditional Monte Carlo) of Chernozhukov, Lee, and Rosen (2013) to construct uniform confidence bands for sieve t-statistics for linear functionals (see their Theorem 5.6).[19] In work that is concurrent with ours, Tao (2014) (Theorem 3.5) extends Belloni et al. (2014)'s results to uniform confidence bands for possibly nonlinear functionals of semi/nonparametric conditional moment restrictions under high-level conditions that are slightly stronger but are similar to those in Chen and Pouzo (2014). Our Theorem B.1 appears to be the first in the literature which establishes consistency of a sieve score bootstrap for uniform inference on general nonlinear functionals of NPIV under low-level conditions.

B.1 Monte Carlo

We now evaluate the performance of our limit theory for uniform confidence bands for h_0. Using the MC design described in Section 3.3, we generate 1000 samples of length 1000 and implement our procedure using B-spline and Legendre polynomial bases as described in Section 3.3. We use a data-driven approach to choose the sieve dimension, taking $\widehat{J}_{\max}$ as described in Section 3.2. For each simulation, we calculate the 90%, 95%, and 99% uniform confidence bands for h_0 over the full support $[0.05, 0.95]$ with 1000 bootstrap replications for each simulation. We draw the innovations for the sieve score bootstrap from the two-point distribution of Mammen (1993). We then calculate the MC coverage probabilities of our uniform confidence bands.

Figure 1 displays the estimate $\widehat{h}$, the structural function h_0, and 90%, 95% and 99% uniform confidence bands for h_0 for a representative sample. Figure 2 displays the estimated structural function and confidence bands together with a scatterplot of the sample (X_i, Y_i) data.

The results of this MC experiment are presented in Table 5. Comparing the MC coverage probabilities with their nominal values, it is clear that the uniform confidence bands for the linear design are

[17] Horowitz and Lee (2012) interpolate h_0 at finitely many grid points with grid size going to zero slowly and prove bootstrap consistency in the case in which the number of interpolation points is finite and fixed.

[18] The assumptions on the moments of u_i and growth conditions on J in Remark B.3 are very similar to those in Horowitz and Lee (2012).

[19] Belloni et al. (2014) also derive a weighted bootstrap uniform Gaussian approximation for linear functionals of series LS when the variance is known (see their Theorem 4.5).

slightly too conservative. However, the uniform confidence bands for the nonlinear design have MC and nominal converge probabilities much closer, with the exception of the quartic B-spline basis. Coverage probabilities of the bands formed using Legendre polynomial bases are particularly good in the nonlinear case.

		$K(J) = J$			$K(J) = 2J$		
r_J	r_K	90% CI	95% CI	99% CI	90% CI	95% CI	99% CI
		Design 1: Linear h_0					
4	4	0.933	0.966	0.996	0.944	0.971	0.994
4	5	0.937	0.975	0.995	0.937	0.963	0.994
5	5	0.961	0.983	0.997	0.959	0.985	0.997
Leg	Leg	0.937	0.964	0.997	0.928	0.959	0.989
		Design 2: Nonlinear h_0					
4	4	0.884	0.945	0.987	0.912	0.956	0.989
4	5	0.894	0.946	0.987	0.906	0.951	0.987
5	5	0.956	0.978	0.995	0.951	0.979	0.996
Leg	Leg	0.901	0.952	0.988	0.906	0.948	0.989

Table 5: MC coverage probabilities of uniform confidence bands for h_0. Results are presented for B-spline bases for Ψ_J and B_K of orders r_J and r_K and Legendre polynomial bases, with two different rules for $K(J)$.

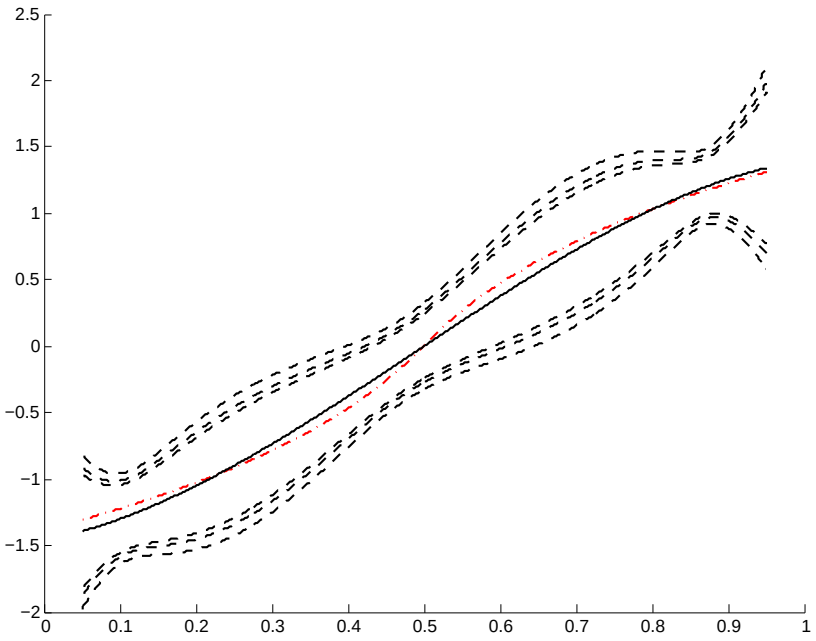

Figure 1: 90%, 95% and 99% uniform confidence bands for h_0 (dashed lines; innermost are 90%, outermost are 99%), estimate $\widehat{h}$ (solid line), and true structural function h_0 (dot-dashed line) for the nonlinear design.

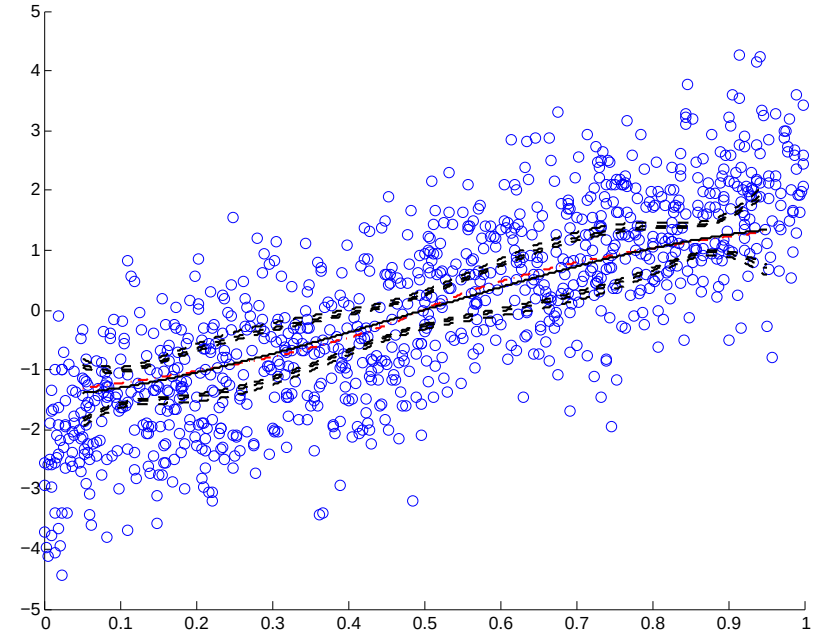

Figure 2: 90%, 95% and 99% uniform confidence bands for h_0 (dashed lines), estimate $\widehat{h}$ (solid line), and true structural function h_0 (dot-dashed line), with (X_i, Y_i) data (circles) for the nonlinear design.

References

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Andrews, D. W. K. (2011). Examples of l2-complete and boundedly-complete distributions. *Cowles Foundation Discussion Paper No. 1801*.

Andrews, D. W. K. and Y. Sun (2004). Adaptive local polynomial whittle estimation of long-range dependence. *Econometrica 72*(2), 569–614.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2014). On the asymptotic theory for least squares series: Pointwise and uniform results. Preprint, arXiv:1212.0442v3 [stat.ME].

Blundell, R., X. Chen, and D. Kristensen (2007). Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica 75*(6), 1613–1669.

Blundell, R., A. Duncan, and C. Meghir (1998). Estimating labor supply responses using tax reforms. *Econometrica 66*(4), pp. 827–861.

Blundell, R., J. L. Horowitz, and M. Parey (2012). Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quantitative Economics 3*(1), 29–51.

Blundell, R., J. L. Horowitz, and M. Parey (2013). Nonparametric estimation of a heterogeneous demand function under the slutsky inequality restriction. Cemmap working paper cwp54/13.

Blundell, R., T. MaCurdy, and C. Meghir (2007). Chapter 69 labor supply models: Unobserved heterogeneity, nonparticipation and dynamics. Volume 6, Part A of *Handbook of Econometrics*, pp. 4667 – 4775. Elsevier.

Breunig, C. and J. Johannes (2013). Adaptive estimation of functionals in nonparametric instrumental regression. Preprint, Universität Mannheim.

Carrasco, M., J.-P. Florens, and E. Renault (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In J. J. Heckman and E. E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6, Part B, Chapter 77, pp. 5633–5751. Elsevier.

Centorrino, S. (2014). Data driven selection of the regularization parameter in additive nonparametric instrumental regressions. Working paper, Stony Brook.

Chen, X. and T. M. Christensen (2013). Optimal uniform convergence rates for sieve nonparametric instrumental variables regression. Cowles Foundation Discussion Paper CFDP1923, Cemmap working paper CWP56/13 and arXiv preprint arXiv:1311.0412 [math.ST].

Chen, X. and T. M. Christensen (2014). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics (forthcoming)*.

Chen, X. and D. Pouzo (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica 80*(1), 277–321.

Chen, X. and D. Pouzo (2014). Sieve Wald and QLR inferences on semi/nonparametric conditional moment models. *Econometrica (forthcoming)*.

Chen, X. and M. Reiss (2011). On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory 27*(3), 497–521.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013). Intersection bounds: Estimation and inference. *Econometrica 81*(2), 667–737.

Darolles, S., Y. Fan, J.-P. Florens, and E. Renault (2011). Nonparametric instrumental regression. *Econometrica 79*(5), 1541–1565.

DeVore, R. A. and G. G. Lorentz (1993). *Constructive Approximation*. Grundlehren der mathematischen Wissenschaften 303. Springer-Verlag, Berlin.

Florens, J.-P. and A. Simoni (2012). Nonparametric estimation of an instrumental variables regression: a quasi-bayesian approach based on regularized posterior. *Journal of Econometrics 170*, 458–475.

Gagliardini, P. and O. Scaillet (2012). Tikhonov regularization for nonparametric instrumental variable estimators. *Journal of Econometrics 167*(1), 61–75.

Gautier, E. and E. LePennec (2011). Adaptive estimation in the nonparametric random coefficients binary choice model by needlet thresholding. Preprint, arXiv:1106.3503v1 [math.ST].

Hall, P. and J. L. Horowitz (2005). Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics 33*(6), 2904–2929.

Hausman, J. A. (1981). Exact consumer's surplus and deadweight loss. *The American Economic Review 71*(4), 662–676.

Hausman, J. A. and W. K. Newey (1995). Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica 63*(6), pp. 1445–1476.

Hoffmann, M. and M. Reiss (2008). Nonlinear estimation for linear inverse problems with error in the operator. *The Annals of Statistics 36*(1), 310–336.

Horowitz, J. L. (2011). Applied nonparametric instrumental variables estimation. *Econometrica 79*(2), 347–394.

Horowitz, J. L. (2014). Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter. *Journal of Econometrics 180*, 158–173.

Horowitz, J. L. and S. Lee (2012). Uniform confidence bands for functions estimated nonparametrically with instrumental variables. *Journal of Econometrics 168*, 175–188.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics 26*(1), 242–272.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics 31*(5), 1600–1635.

Johannes, J. and M. Schwarz (2013). Adaptive gaussian inverse regression with partially unknown operator. *Communications in Statistics—Theory and Methods, 42*, 1343–1362.

Kato, K. (2013). Quasi-bayesian analysis of nonparametric instrumental variables models. *The Annals of Statistics 41*(5), 2359–2390.

Lepskii, O. V. (1990). On a problem of adaptive estimation in gaussian white noise. *Theory of Probability and its Applications 35*(3), 454–466.

Li, K.-C. (1987). Asymptotic optimality for cp, cl, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, 958–975.

Liao, Y. and W. Jiang (2011). Posterior consistency of nonparametric conditional moment restricted models. *The Annals of Statistics 39*(6), 3003–3031.

Liu, C.-A. and J. Tao (2014). Model selection and model averaging in nonparametric instrumental variables models. Working paper, National University of Singapore and University of Wisconsin-Madison.

Loubes, J.-M. and C. Marteau (2012). Adaptive estimation for an inverse regression model with unknown operator. *Statistics & Risk Modeling 29*(3), 215–242.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics 21*(1), 255–285.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Newey, W. K. (2013). Nonparametric instrumental variables estimation. *American Economic Review: Papers and Proceedings 103*(3), 550–56.

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Schmalensee, R. and T. M. Stoker (1999). Household gasoline demand in the united states. *Econometrica 67*(3), 645–662.

Sueishi, N. (2012). Model selection criterion for instrumental variable models. Working paper, Kyoto University.

Tao, J. (2014). Inference for point and partially identified semi-nonparametric conditional moment models. Working paper, University of Wisconsin-Madison.

Triebel, H. (2006). *Theory of Function Spaces III*. Birkhäuser, Basel.

Vanhems, A. (2010). Non-parametric estimation of exact consumer surplus with endogeneity in price. *Econometrics Journal 13*(3), S80–S98.

Yatchew, A. and J. A. No (2001). Household gasoline demand in Canada. *Econometrica 69*(6), 1697–1709.