

Optimal Uniform Convergence Rates for Sieve Nonparametric Instrumental Variables Regression*

Xiaohong Chen[†] and Timothy M. Christensen[‡]

First version January 2012; Revised August 2013

Abstract

We study the problem of nonparametric regression when the regressor is endogenous, which is an important nonparametric instrumental variables (NPIV) regression in econometrics and a difficult ill-posed inverse problem with unknown operator in statistics. We first establish a general upper bound on the sup-norm (uniform) convergence rate of a sieve estimator, allowing for endogenous regressors and weakly dependent data. This result leads to the optimal sup-norm convergence rates for spline and wavelet least squares regression estimators under weakly dependent data and heavy-tailed error terms. This upper bound also yields the sup-norm convergence rates for sieve NPIV estimators under i.i.d. data: the rates coincide with the known optimal L^2 -norm rates for severely ill-posed problems, and are power of $\log(n)$ slower than the optimal L^2 -norm rates for mildly ill-posed problems. We then establish the minimax risk lower bound in sup-norm loss, which coincides with our upper bounds on sup-norm rates for the spline and wavelet sieve NPIV estimators. This sup-norm rate optimality provides another justification for the wide application of sieve NPIV estimators. Useful results on weakly-dependent random matrices are also provided.

JEL Classification: C13, C14, C32

Key words and phrases: Nonparametric instrumental variables; Statistical ill-posed inverse problems; Optimal uniform convergence rates; Weak dependence; Random matrices; Splines; Wavelets

*Support from the Cowles Foundation is gratefully acknowledged. We thank conference participants of SETA2013 in Seoul and AMES2013 in Singapore for useful comments. Any errors are the responsibility of the authors.

[†]Cowles Foundation for Research in Economics, Yale University: xiaohong.chen@yale.edu

[‡]Department of Economics, Yale University: timothy.christensen@yale.edu

1 Introduction

In economics and other social sciences one frequently encounters the relation

$$Y_{1i} = h_0(Y_{2i}) + \epsilon_i \tag{1}$$

where Y_{1i} is a response variable, Y_{2i} is a predictor variable, h_0 is an unknown structural function of interest, and ϵ_i is an error term. However, a latent external mechanism may “determine” or “cause” Y_{1i} and Y_{2i} simultaneously, in which case the conditional mean restriction $E[\epsilon_i|Y_{2i}] = 0$ fails and Y_{2i} is said to be *endogenous*.¹ When the regressor Y_{2i} is endogenous one cannot use standard nonparametric regression techniques to consistently estimate h_0 . In this instance one typically assumes that there exists a vector of *instrumental variables* X_i such that $E[\epsilon_i|X_i] = 0$ and for which there is a nondegenerate relationship between X_i and Y_{2i} . Such a setting permits estimation of h_0 using nonparametric instrumental variables (NPIV) techniques based on a sample $\{(X_i, Y_{1i}, Y_{2i})\}_{i=1}^n$. In this paper we assume that the data is strictly stationary in that (X_i, Y_{1i}, Y_{2i}) has the same (unknown) distribution F_{X, Y_1, Y_2} as that of (X, Y_1, Y_2) for all i .²

NPIV estimation has been the subject of much research in recent years, both because of its practical importance to applied economics and its prominent role in the literature on linear ill-posed inverse problems with unknown operators. In many economic applications the joint distribution F_{X, Y_2} of X_i and Y_{2i} is unknown but is assumed to have a continuous density. Therefore the conditional expectation operator $Th(\cdot) = E[h(Y_{2i})|X_i = \cdot]$ is typically unknown but compact. Model (1) with $E[\epsilon_i|X_i] = 0$ can be equivalently written as

$$\begin{aligned} Y_{1i} &= Th_0(X_i) + u_i \\ E[u_i|X_i] &= 0 \end{aligned} \tag{2}$$

where $u_i = h_0(Y_{2i}) - Th_0(X_i) + \epsilon_i$. Model (2) is called the reduced-form NPIV model if T is assumed to be unknown and the nonparametric indirect regression (NPIR) model if T is assumed to be known. Let $\widehat{E}[Y_1|X = \cdot]$ be a consistent estimator of $E[Y_1|X = \cdot]$. Regardless of whether the compact operator T is unknown or known, nonparametric recovery of h_0 by inversion of the conditional expectation operator T on the left-hand side of the Fredholm equation of the first kind

$$Th(\cdot) = \widehat{E}[Y_1|X = \cdot] \tag{3}$$

¹In a canonical example of this relation, Y_{1i} may be the hourly wage of person i and Y_{2i} may include the education level of person i . The latent ability of person i affects both Y_{1i} and Y_{2i} . See Blundell and Powell (2003) for other examples and discussions of endogeneity in semi/nonparametric regression models.

²The subscript i denotes either the individual i in a cross-sectional sample or the time period i in a time-series sample. Since the sample is strictly stationary we sometimes drop the subscript i without confusion.

leads to an ill-posed inverse problem (see, e.g., Kress (1999)). Consequently, some form of regularization is required for consistent nonparametric estimation of h_0 . In the literature there are several popular methods of NPIV estimation, including but not limited to (1) finite-dimensional sieve minimum distance estimators (Newey and Powell, 2003; Ai and Chen, 2003; Blundell, Chen, and Kristensen, 2007); (2) kernel-based Tikhonov regularization estimators (Hall and Horowitz, 2005; Darolles, Fan, Florens, and Renault, 2011; Gagliardini and Scaillet, 2012) and their Bayesian version (Florens and Simoni, 2012); (3) orthogonal series Tikhonov regularization estimators (Hall and Horowitz, 2005); (4) orthogonal series Galerkin-type estimators (Horowitz, 2011); (5) general penalized sieve minimum distance estimators (Chen and Pouzo, 2012) and their Bayesian version (Liao and Jiang, 2011). See Horowitz (2011) for a recent review and additional references.

To the best of our knowledge, all the existing works on convergence rates for various NPIV estimators have only studied L^2 -norm convergence rates. In particular, Hall and Horowitz (2005) are the first to establish the minimax risk lower bound in L^2 -norm loss for a class of mildly ill-posed NPIV models, and show that their estimators attain the lower bound. Chen and Reiss (2011) derive the minimax risk lower bound in L^2 -norm loss for a large class of NPIV models that could be mildly or severely ill-posed, and show that the sieve minimum distance estimator of Blundell, Chen, and Kristensen (2007) achieves the lower bound. Subsequently, some other NPIV estimators listed above have also been shown to achieve the optimal L^2 -norm convergence rates. As yet there are no published results on sup-norm (uniform) convergence rates for any NPIV estimators, nor results on what are the minimax risk lower bounds in sup-norm loss for any class of NPIV models.

Sup-norm convergence rates for any estimators of h_0 are important for constructing uniform confidence bands for the unknown h_0 in NPIV models and for conducting inference on nonlinear functionals of h_0 , but are currently missing. In this paper we study the uniform convergence properties of the sieve minimum distance estimator of h_0 for the NPIV model, which is a nonparametric series two-stage least squares regression estimator (Newey and Powell, 2003; Ai and Chen, 2003; Blundell, Chen, and Kristensen, 2007). We focus on this estimator because it is easy to compute and has been used in empirical work in demand analysis (Blundell, Chen, and Kristensen, 2007; Chen and Pouzo, 2009), asset pricing (Chen and Ludvigson, 2009), and other applied fields in economics. Also, this class of estimators is known to achieve the optimal L^2 -norm convergence rates for both mildly and severely ill-posed NPIV models.

We first establish a general upper bound (Theorem 2.1) on the uniform convergence rate of a sieve estimator, allowing for endogenous regressors and weakly dependent data. To provide sharp bounds on the sieve approximation error or “bias term” we extend the proof strategy of Huang (2003) for sieve nonparametric least squares (LS) regression to the sieve NPIV estimator. Together, these tools yield sup-norm convergence rates for the spline and wavelet sieve NPIV estimators

under i.i.d. data. Under conditions similar to those for the L^2 -norm convergence rates for the sieve NPIV estimators, our sup-norm convergence rates coincide with the known optimal L^2 -norm rates for severely ill-posed problems, and are power of $\log(n)$ slower than the optimal L^2 -norm rates for mildly ill-posed problems. We then establish the minimax risk lower bound in sup-norm loss for h_0 in a NPIR model (i.e., (2) with a known compact T) uniformly over Hölder balls, which in turn provides a lower bound in sup-norm loss for h_0 in a NPIV model uniformly over Hölder balls. The lower bound is shown to coincide with our sup-norm convergence rates for the spline and wavelet sieve NPIV estimators.

To establish the general upper bound, we first derive a new exponential inequality for sums of weakly dependent random matrices in Section 5. This allows us to weaken conditions under which the optimal uniform convergence rates can be obtained. As an indication of the sharpness of our general upper bound result, we show that it leads to the optimal uniform convergence rates for spline and wavelet LS regression estimators with weakly dependent data and heavy-tailed error terms. Precisely, for beta-mixing dependent data and finite $(2 + \delta)$ -th moment error term (for $\delta \in (0, 2)$), we show that the spline and wavelet nonparametric LS regression estimators attain the minimax risk lower bound in sup-norm loss of Stone (1982). This result should be very useful to the literature on nonparametric estimation with financial time series.

The NPIV model falls within the class of statistical linear ill-posed inverse problems with *unknown* operators and additive noise. There is a vast literature on statistical linear ill-posed inverse problems with known operators and additive noise. Some recent references include but are not limited to Cavalier, Golubev, Picard, and Tsybakov (2002), Cohen, Hoffmann, and Reiss (2004) and Cavalier (2008), of which density deconvolution is an important and extensively-studied problem (see, e.g., Carroll and Hall (1988); Zhang (1990); Fan (1991); Hall and Meister (2007); Lounici and Nickl (2011)). There are also papers on statistical linear ill-posed inverse problems with pseudo-unknown operators (i.e., known eigenfunctions but unknown singular values) (see, e.g., Cavalier and Hengartner (2005), Loubes and Marteau (2012)). Related papers that allow for an unknown linear operator but assume the existence of an estimator of the operator (with rate) include Efromovich and Koltchinskii (2001), Hoffmann and Reiss (2008) and others. To the best of our knowledge, most of the published works in the statistical literature on linear ill-posed inverse problems also focus on the rate optimality in L^2 -norm loss, except that of Lounici and Nickl (2011) which recently establishes the optimal sup-norm convergence rate for a wavelet density deconvolution estimator. Therefore, our minimax risk lower bounds in sup-norm loss for the NPIR and NPIV models also contribute to the large literature on statistical ill-posed inverse problems.

The rest of the paper is organized as follows. Section 2 outlines the model and presents a general upper bound on the uniform convergence rates for a sieve estimator. Section 3 establishes

the optimal uniform convergence rates for the sieve NPIV estimators, allowing for both mildly and severely ill-posed inverse problems. Section 4 derives the optimal uniform convergence rates for the sieve nonparametric (least squares) regression, allowing for dependent data. Section 5 provides useful exponential inequalities for sums of random matrices, and the reinterpretation of equivalence of the theoretical and empirical L^2 norms as a criterion regarding convergence of a random matrix. The appendix contains a brief review of the spline and wavelet sieve spaces, proofs of all the results in the main text, and supplementary results.

Notation: $\|\cdot\|$ denotes the Euclidean norm when applied to vectors and the matrix spectral norm (largest singular value) when applied to matrices. For a random variable Z let $L^q(Z)$ denote the spaces of (equivalence classes of) measurable functions of z with finite q -th moment if $1 \leq q < \infty$ and let $\|\cdot\|_{L^q(Z)}$ denote the $L^q(Z)$ norm. Let $L^\infty(Z)$ denote the space of measurable functions of z with finite sup norm $\|\cdot\|_\infty$. If A is a square matrix, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its smallest and largest eigenvalues, respectively, and A^- denotes its Moore-Penrose generalized inverse. If $\{a_n : n \geq 1\}$ and $\{b_n : n \geq 1\}$ are two sequences of non-negative numbers, $a_n \lesssim b_n$ means there exists a finite positive C such that $a_n \leq Cb_n$ for all n sufficiently large, and $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. $\#(\mathcal{S})$ denotes the cardinality of a set \mathcal{S} of finitely many elements. Let $\text{BSpl}(K, [0, 1]^d, \gamma)$ and $\text{Wav}(K, [0, 1]^d, \gamma)$ denote tensor-product B-spline (with smoothness γ) and wavelet (with regularity γ) sieve spaces of dimension K on $[0, 1]^d$ (see Appendix A for details on construction of these spaces).

2 Uniform convergence rates for sieve NPIV estimators

We begin by considering the NPIV model

$$\begin{aligned} Y_{1i} &= h_0(Y_{2i}) + \epsilon_i \\ E[\epsilon_i | X_i] &= 0 \end{aligned} \tag{4}$$

where $Y_1 \in \mathbb{R}$ is a response variable, Y_2 is an endogenous regressor with support $\mathcal{Y}_2 \subset \mathbb{R}^d$ and X is a vector of conditioning variables (also called instruments) with support $\mathcal{X} \subset \mathbb{R}^{d_x}$. The object of interest is the unknown structural function $h_0 : \mathcal{Y}_2 \rightarrow \mathbb{R}$ which belongs to some infinite-dimensional parameter space $\mathcal{H} \subset L^2(\mathcal{Y}_2)$. It is assumed hereafter that h_0 is identified uniquely by the conditional moment restriction (4). See Newey and Powell (2003), Blundell, Chen, and Kristensen (2007), Darolles, Fan, Florens, and Renault (2011), Andrews (2011), D'Haultfoeuille (2011), Chen, Chernozhukov, Lee, and Newey (2013) and references therein for sufficient conditions for identification.

2.1 Sieve NPIV estimators

The sieve NPIV estimator due to Newey and Powell (2003), Ai and Chen (2003), and Blundell, Chen, and Kristensen (2007) is a nonparametric series two-stage least squares estimator. Let the sieve spaces $\{\Psi_J : J \geq 1\} \subseteq L^2(Y_2)$ and $\{B_K : K \geq 1\} \subseteq L^2(X)$ be sequences of subspaces of dimension J and K spanned by sieve basis functions such that Ψ_J and B_K become dense in $\mathcal{H} \subseteq L^2(Y_2)$ and $L^2(X)$ as $J, K \rightarrow \infty$. For given J and K , let $\{\psi_{J1}, \dots, \psi_{JJ}\}$ and $\{b_{K1}, \dots, b_{KK}\}$ be sets of sieve basis functions whose closed linear span generates Ψ_J and B_K respectively. We consider sieve spaces generated by spline, wavelet or other Riesz basis functions that have nice approximation properties (see Section 3 for details).

In the first stage, the conditional moment function $m(x, h) : \mathcal{X} \times \mathcal{H} \rightarrow \mathbb{R}$ given by

$$m(x, h) = E[Y_1 - h(Y_2)|X = x] \quad (5)$$

is estimated using the series (least squares) regression estimator

$$\hat{m}(x, h) = \sum_{i=1}^n b^K(x)'(B'B)^{-1}b^K(X_i)(Y_{1i} - h(Y_{2i})) \quad (6)$$

where

$$\begin{aligned} b^K(x) &= (b_{K1}(x), \dots, b_{KK}(x))' \\ B &= (b^K(X_1), \dots, b^K(X_n))'. \end{aligned} \quad (7)$$

The sieve NPIV estimator \hat{h} is then defined as the solution to the second-stage minimization problem

$$\hat{h} = \arg \min_{h \in \Psi_J} \frac{1}{n} \sum_{i=1}^n \hat{m}(X_i, h)^2 \quad (8)$$

which may be solved in closed form to give

$$\hat{h}(y_2) = \psi^J(y_2)'[\Psi'B(B'B)^{-1}B'\Psi]^{-1}\Psi'B(B'B)^{-1}B'Y \quad (9)$$

where

$$\begin{aligned} \psi^J(y_2) &= (\psi_{J1}(y_2), \dots, \psi_{JJ}(y_2))' \\ \Psi &= (\psi^J(Y_{21}), \dots, \psi^J(Y_{2n}))' \\ Y &= (Y_{11}, \dots, Y_{1n})'. \end{aligned} \quad (10)$$

Under mild regularity conditions (see Newey and Powell (2003), Blundell, Chen, and Kristensen (2007) and Chen and Pouzo (2012)), \hat{h} is a consistent estimator of h_0 (in both $\|\cdot\|_{L^2(Y_2)}$ and $\|\cdot\|_\infty$ norms) as $n, J, K \rightarrow \infty$, provided $J \leq K$ and J increases appropriately slowly so as to regularize the ill-posed inverse problem.³ We note that the modified sieve estimator (or orthogonal series

³Here we have used K to denote the “smoothing parameter” (i.e. the dimension of the sieve space used to estimate the conditional moments in (6)) and J to denote the “regularization parameter” (i.e. the dimension of the sieve space used to approximate the unknown h_0). Note that Chen and Reiss (2011) use J and m , Blundell, Chen, and Kristensen (2007) and Chen and Pouzo (2012) use J and k to denote the smoothing and regularization parameters, respectively.

Galerkin-type estimator) of Horowitz (2011) corresponds to the sieve NPIV estimator with $J = K$ and $\psi^J(\cdot) = b^K(\cdot)$ being orthonormal basis in $L^2(\text{Lebesgue})$.

2.2 A general upper bound on uniform convergence rates for sieve estimators

We first present a general calculation for sup-norm convergence which will be used to obtain uniform convergence rates for both the sieve NPIV and the sieve LS estimators below.

As the sieve estimators are invariant to an invertible transformation of the sieve basis functions, we re-normalize the sieve spaces B_K and Ψ_J so that $\{\tilde{b}_{K1}, \dots, \tilde{b}_{KK}\}$ and $\{\tilde{\psi}_{J1}, \dots, \tilde{\psi}_{JJ}\}$ form orthonormal bases for B_K and Ψ_J . This is achieved by setting $\tilde{b}^K(x) = E[b^K(X)b^K(X)']^{-1/2}b^K(x)$ where $^{-1/2}$ denotes the inverse of the positive-definite matrix square root (which exists under Assumption 4(ii) below), with $\tilde{\psi}^J$ similarly defined. Let

$$\begin{aligned}\tilde{B} &= (\tilde{b}^K(X_1), \dots, \tilde{b}^K(X_n))' \\ \tilde{\Psi} &= (\tilde{\psi}^J(Y_{21}), \dots, \tilde{\psi}^J(Y_{2n}))'\end{aligned}\tag{11}$$

and define the $J \times K$ matrices

$$\begin{aligned}S &= E[\tilde{\psi}^J(Y_2)\tilde{b}^K(X)'] \\ \hat{S} &= \tilde{\Psi}'\tilde{B}/n.\end{aligned}\tag{12}$$

Let $\sigma_{JK}^2 = \lambda_{\min}(SS')$. For each $h \in \Psi_J$ define

$$\Pi_K Th(\cdot) = \tilde{b}^K(x)'E[\tilde{b}^K(X)(Th)(X)] = \tilde{b}^K(\cdot)'E[\tilde{b}^K(X)h(Y_2)]\tag{13}$$

which is the $L^2(X)$ orthogonal projection of $Th(\cdot)$ onto B_K . The variational characterization of singular values gives

$$\sigma_{JK} = \inf_{h \in \Psi_J: \|h\|_{L^2(Y_2)}=1} \|\Pi_K Th\|_{L^2(X)} \leq 1.\tag{14}$$

Finally, define P_n as the second-stage empirical projection operator onto the sieve space Ψ_J after projecting onto the instrument space B_K , viz.

$$P_n h_0(y_2) = \tilde{\psi}^J(y_2)[\hat{S}(\tilde{B}'\tilde{B}/n)^{-}\hat{S}']^{-}\hat{S}(\tilde{B}'\tilde{B}/n)^{-}\tilde{B}'H_0/n\tag{15}$$

where $H_0 = (h_0(Y_{21}), \dots, h_0(Y_{2n}))'$.

We first decompose the sup-norm error as

$$\|h_0 - \hat{h}\|_{\infty} \leq \|h_0 - P_n h_0\|_{\infty} + \|P_n h_0 - \hat{h}\|_{\infty}\tag{16}$$

and calculate the uniform convergence rate for the ‘‘variance term’’ $\|\hat{h} - P_n h_0\|_{\infty}$ in this section. Control of the ‘‘bias term’’ $\|h_0 - P_n h_0\|_{\infty}$ is left to the subsequent sections, which will be dealt with under additional regularity conditions for the NPIV model and the LS regression model separately.

Let $Z_i = (X_i, Y_{1i}, Y_{2i})$ and $\mathcal{F}_{i-1} = \sigma(X_i, X_{i-1}, \epsilon_{i-1}, X_{i-2}, \epsilon_{i-2}, \dots)$.

Assumption 1 (i) $\{Z_i\}_{i=-\infty}^{\infty}$ is strictly stationary, (ii) X has support $\mathcal{X} = [0, 1]^d$ and Y_2 has support $\mathcal{Y}_2 = [0, 1]^d$, (iii) the distributions of X and Y_2 have density (with respect to Lebesgue measure) which is uniformly bounded away from zero and infinity over \mathcal{X} and \mathcal{Y}_2 respectively.

The results stated in this section do not actually require that $\dim(X) = \dim(Y_2)$. However, most published papers on NPIV models assume $\dim(X) = \dim(Y_2) = d$ and so we follow this convention in Assumption 1(ii).

Assumption 2 (i) $(\epsilon_i, \mathcal{F}_{i-1})_{i=-\infty}^{\infty}$ is a strictly stationary martingale difference sequence, (ii) the conditional second moment $E[\epsilon_i^2 | \mathcal{F}_{i-1}]$ is uniformly bounded away from zero and infinity, (iii) $E[|\epsilon_i|^{2+\delta}] < \infty$ for some $\delta > 0$.

Assumption 3 (i) Sieve basis $\psi^J(\cdot)$ is Hölder continuous with smoothness $\gamma > p$ and $\sup_{y_2 \in \mathcal{Y}_2} \|\psi^J(y_2)\| \lesssim \sqrt{J}$, (ii) $\lambda_{\min}(E[\psi^J(Y_2)\psi^J(Y_2)']) \geq \underline{\lambda} > 0$ for all $J \geq 1$.

In what follows, $p > 0$ indicates the smoothness of the function $h_0(\cdot)$ (see Assumption 5 in Section 3).

Assumption 4 (i) Sieve basis $b^K(\cdot)$ is Hölder continuous with smoothness $\gamma_x \geq \gamma > p$ and $\sup_{x \in \mathcal{X}} \|b^K(x)\| \lesssim \sqrt{K}$, (ii) $\lambda_{\min}(E[b^K(X)b^K(X)']) \geq \underline{\lambda} > 0$ for all $K \geq 1$.

The preceding assumptions on the data generating process trivially nest i.i.d. sequences but also allow for quite general weakly-dependent data. In an i.i.d. setting, Assumption 2(ii) reduces to requiring that $E[\epsilon_i^2 | X_i = x]$ be bounded uniformly from zero and infinity which is standard (see, e.g., Newey (1997); Hall and Horowitz (2005)). The value of δ in Assumption 2(iii) depends on the context. For example, $\delta \geq d/p$ will be shown to be sufficient to attain the optimal sup-norm convergence rates for series LS regression in Section 4, whereas lower values of δ suffice to attain the optimal sup-norm convergence rates for the sieve NPIV estimator in Section 3. Rectangular support and bounded densities of the endogenous regressor and instrument are assumed in Hall and Horowitz (2005). Assumptions 3(i) and 4(i) are satisfied by many widely used sieve bases such as spline, wavelet and cosine sieves, but they rule out polynomial and power series sieves (see, e.g., Newey (1997); Huang (1998)). The instruments sieve basis $b^K(\cdot)$ is used to approximate the conditional expectation operator $Th = E[h(Y_2)|X = \cdot]$, which is a smoothing operator. Thus Assumption 4(i) assumes that the sieve basis $b^K(\cdot)$ (for Th) is smoother than that of the sieve basis $\psi^J(\cdot)$ (for h).

In the next theorem, our upper bound on the “variance term” $\|\widehat{h} - P_n h_0\|_{\infty}$ holds under general weak dependence as captured by Condition (ii) on the convergence of the random matrices $\widetilde{B}'\widetilde{B}/n - I_K$ and $\widehat{S} - S$.

Theorem 2.1 *Let Assumptions 1, 2, 3 and 4 hold. If $\sigma_{JK} > 0$ then:*

$$\|h_0 - \hat{h}\|_\infty \leq \|h_0 - P_n h_0\|_\infty + O_p\left(\sigma_{JK}^{-1} \sqrt{K(\log n)/n}\right)$$

provided $n, J, K \rightarrow \infty$ and

$$(i) \quad J \leq K, \quad K \lesssim (n/\log n)^{\delta/(2+\delta)}, \quad \text{and} \quad \sigma_{JK}^{-1} \sqrt{K(\log n)/n} \lesssim 1$$

$$(ii) \quad \sigma_{JK}^{-1} \left(\|(\tilde{B}'\tilde{B}/n) - I_K\| + \|\hat{S} - S\| \right) = O_p(\sqrt{(\log n)/K}) = o_p(1).$$

The restrictions on J , K and n in Conditions (i) and (ii) merit a brief explanation. The restriction $J \leq K$ merely ensures that the sieve NPIV estimator is well defined. The restriction $K \lesssim (n/\log n)^{\delta/(2+\delta)}$ is used to perform a truncation argument using the existence of $(2 + \delta)$ -th moment of the error terms (see Assumption 2). Condition (ii) ensures that J increases sufficiently slowly that with probability approaching one the minimum eigenvalue of the “denominator” matrix $\Psi' B(B'B)^{-1} B' \Psi/n$ is positive and bounded below by a multiple of σ_{JK}^2 , thereby regularizing the ill-posed inverse problem. It also ensures the error in estimating the matrices $(\tilde{B}'\tilde{B}/n)$ and \hat{S} vanishes sufficiently quickly that it doesn't affect the convergence rate of the estimator.

Remark 2.1 *Section 5 provides very mild low-level sufficient conditions for Condition (ii) to hold under weakly dependent data. In particular, when specializing Corollary 5.1 to i.i.d. data $\{(X_i, Y_{2i})\}_{i=1}^n$ (also see Lemma 5.2), under Assumptions 3 and 4 and $J \leq K$, we have:*

$$\|(\tilde{B}'\tilde{B}/n) - I_K\| = O_p(\sqrt{K(\log K)/n}), \quad \|\hat{S} - S\| = O_p(\sqrt{K(\log K)/n}).$$

3 Optimal uniform convergence rates for sieve NPIV estimators

3.1 Upper bounds on uniform convergence rates for sieve NPIV estimators

We now exploit the specific linear structure of the sieve NPIV estimator to derive uniform convergence rates for the mildly and severely ill-posed cases. Some additional assumptions are required so as to control the “bias term” $\|h_0 - P_n h_0\|_\infty$ and to relate the estimator to the measure of ill-posedness.

p -smooth Hölder class of functions. We first impose a standard smoothness condition on the unknown structural function h_0 to facilitate comparison with Stone (1982)'s minimax risk lower bound in sup-norm loss for a nonparametric regression function. Recall that $\mathcal{Y}_2 = [0, 1]^d$. Deferring definitions to Triebel (2006, 2008), we let $B_{q,q}^p([0, 1]^d)$ denote the Besov space of smoothness p on the domain $[0, 1]^d$ and $\|\cdot\|_{B_{q,q}^p}$ denote the usual Besov norm on this space. Special cases include the Sobolev class of smoothness p , namely $B_{2,2}^p([0, 1]^d)$, and the Hölder-Zygmund class of smoothness p , namely $B_{\infty,\infty}^p([0, 1]^d)$. Let $B(p, L)$ denote a Hölder ball of smoothness p and radius $0 < L < \infty$, i.e. $B(p, L) = \{h \in B_{\infty,\infty}^p([0, 1]^d) : \|h\|_{B_{\infty,\infty}^p} \leq L\}$.

Assumption 5 $h_0 \in \mathcal{H} = B_{\infty, \infty}^p([0, 1]^d)$ for some $p \geq d/2$.

Assumptions 3 and 5 imply that there is $\pi_J h_0 \in \Psi_J$ such that $\|h_0 - \pi_J h_0\|_{\infty} = O(J^{-p/d})$.

Sieve measure of ill-posedness. Let $T : L^q(Y_2) \rightarrow L^q(X)$ denote the conditional expectation operator for $1 \leq q \leq \infty$:

$$Th(x) = E[h(Y_{2i}) | X_i = x]. \quad (17)$$

When Y_2 is endogenous, T is compact under mild conditions on the conditional density of Y_2 given X . For $q' \geq q \geq 1$, we define a measure of ill-posedness (over a sieve space Ψ_J) as

$$\tau_{q, q', J} = \sup_{h \in \Psi_J : \|Th\|_{L^q(X)} \neq 0} \frac{\|h\|_{L^{q'}(Y_2)}}{\|Th\|_{L^q(X)}}. \quad (18)$$

The $\tau_{2, 2, J}$ measure of ill-posedness is clearly related to our earlier definition of σ_{JK} . By definition

$$\sigma_{JK} = \inf_{h \in \Psi_J : \|h\|_{L^2(Y_2)} = 1} \|\Pi_K Th\|_{L^2(X)} \leq \inf_{h \in \Psi_J : \|h\|_{L^2(Y_2)} = 1} \|Th\|_{L^2(X)} = (\tau_{2, 2, J})^{-1}$$

when $J \leq K$. The sieve measures of ill-posedness, $\tau_{2, 2, J}$ and σ_{JK}^{-1} , are clearly non-decreasing in J . In Blundell, Chen, and Kristensen (2007), Horowitz (2011) and Chen and Pouzo (2012), the NPIV model is said to be

- *mildly ill-posed* if $\tau_{2, 2, J} = O(J^{\varsigma/d})$ for some $\varsigma > 0$;
- *severely ill-posed* if $\tau_{2, 2, J} = O(\exp(\frac{1}{2}J^{\varsigma/d}))$ for some $\varsigma > 0$.

These measures of ill-posedness are not exactly the same as (but are related to) the measure of ill-posedness used in Hall and Horowitz (2005) and Cavalier (2008). In the latter papers, it is assumed that the compact operator $T : L^2(Y_2) \rightarrow L^2(X)$ admits a singular value decomposition $\{\mu_k; \phi_{1k}, \phi_{0k}\}_{k=1}^{\infty}$, where $\{\mu_k\}_{k=1}^{\infty}$ are the singular numbers arranged in non-increasing order ($\mu_k \geq \mu_{k+1} \searrow 0$), $\{\phi_{1k}(y_2)\}_{k=1}^{\infty}$ and $\{\phi_{0k}(x)\}_{k=1}^{\infty}$ are eigenfunction (orthonormal) bases for $L^2(Y_2)$ and $L^2(X)$ respectively, and ill-posedness is measured in terms of the rate of decay of the singular values towards zero. Denote T^* as the adjoint operator of T : $\{T^*g\}(Y_2) \equiv E[g(X) | Y_2]$, which maps $L^2(X)$ into $L^2(Y_2)$. Then a compact T implies that T^* , T^*T and TT^* are also compact, and that $T\phi_{1k} = \mu_k\phi_{0k}$ and $T^*\phi_{0k} = \mu_k\phi_{1k}$ for all k . We note that $\|Th\|_{L^2(X)} = \|(T^*T)^{1/2}h\|_{L^2(Y_2)}$ for all $h \in \text{Dom}(T)$. The following lemma provides some relations between these different measures of ill-posedness.

Lemma 3.1 *Let the conditional expectation operator $T : L^2(Y_2) \rightarrow L^2(X)$ be compact and injective. Then: (1) $\sigma_{JK}^{-1} \geq \tau_{2, 2, J} \geq 1/\mu_J$; (2) If the sieve space Ψ_J spans the closed linear subspace (in $L^2(Y_2)$) generated by $\{\phi_{1k} : k = 1, \dots, J\}$, then: $\tau_{2, 2, J} \leq 1/\mu_J$; (3) If, in addition, $J \leq K$ and the sieve space B_K contains the closed linear subspace (in $L^2(X)$) generated by $\{\phi_{0k} : k = 1, \dots, J\}$, then: $\sigma_{JK}^{-1} \leq 1/\mu_J$ and hence $\sigma_{JK}^{-1} = \tau_{2, 2, J} = 1/\mu_J$.*

Lemma 3.1 parts (1) and (2) is Lemma 1 of Blundell, Chen, and Kristensen (2007), while Lemma 3.1 part (3) is proved in the Appendix. We next present a sufficient condition to bound the sieve measures of ill-posedness σ_{JK}^{-1} and $\tau_{2,2,J}$.

Assumption 6 (sieve reverse link condition) *There is a continuous increasing function $\varphi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that: (a) $\|Th\|_{L^2(X)}^2 \gtrsim \sum_{j=1}^J \varphi(j^{-2/d}) |E[h(Y_2)\tilde{\psi}_{Jj}(Y_2)]|^2$ for all $h \in \Psi_J$; or (b) $\|\Pi_K Th\|_{L^2(X)}^2 \gtrsim \sum_{j=1}^J \varphi(j^{-2/d}) |E[h(Y_2)\tilde{\psi}_{Jj}(Y_2)]|^2$ for all $h \in \Psi_J$*

It is clear that Assumption 6(b) implies Assumption 6(a). Assumption 6(a) is the so-called “sieve reverse link condition” used in Chen and Pouzo (2012), which is weaker than the “reverse link condition” imposed in Chen and Reiss (2011) and others in the ill-posed inverse literature: $\|Th\|_{L^2(X)}^2 \gtrsim \sum_{j=1}^\infty \varphi(j^{-2/d}) |E[h(Y_2)\tilde{\psi}_{Jj}(Y_2)]|^2$ for all $h \in B(p, L)$. We immediately have the following bounds:

Remark 3.1 (1) *Assumption 6(a) implies that $\tau_{2,2,J} \lesssim (\varphi(J^{-2/d}))^{-1/2}$. (2) *Assumption 6(b) implies that $\tau_{2,2,J} \leq \sigma_{JK}^{-1} \lesssim (\varphi(J^{-2/d}))^{-1/2}$.**

Given Remark 3.1, in this paper we could call a NPIV model

- *mildly ill-posed* if $\sigma_{JK}^{-1} = O(J^{\zeta/d})$ or $\varphi(t) = t^\zeta$ for some $\zeta > 0$;
- *severely ill-posed* if $\sigma_{JK}^{-1} = O(\exp(\frac{1}{2}J^{\zeta/d}))$ or $\varphi(t) = \exp(-t^{-\zeta/2})$ for some $\zeta > 0$.

Define

$$\sigma_{\infty,JK} = \inf_{h \in \Psi_J: \|h\|_\infty=1} \|\Pi_K Th\|_\infty \leq (\tau_{\infty,\infty,J})^{-1}. \quad (19)$$

Assumption 7 (i) *The conditional expectation operator $T : L^q(Y_2) \rightarrow L^q(X)$ is compact and injective for $q = 2$ and $q = \infty$, (ii) $\sigma_{\infty,JK}^{-1} \|\Pi_K T(h_0 - \pi_J h_0)\|_\infty \lesssim \|h_0 - \pi_J h_0\|_\infty$.*

Assumption 7(ii) is a sup-norm analogue of the so-called “stability condition” imposed in the ill-posed inverse regression literature, such as Assumption 6 of Blundell, Chen, and Kristensen (2007) and Assumption 5.2(ii) of Chen and Pouzo (2012).

To control the “bias term” $\|P_n h_0 - h_0\|_\infty$, we will use spline or wavelet sieves in Assumptions 3 and 4 so that we can make use of sharp bounds on the approximation error due to Huang (2003).⁴ Control of the “bias term” $\|P_n h_0 - h_0\|_\infty$ is more involved in the sieve NPIV context than the sieve nonparametric LS regression context. In particular, control of this term makes use of an additional argument using exponential inequalities. To simplify presentation, the next theorem just presents the uniform convergence rate for sieve NPIV estimators under i.i.d. data.

⁴The key property of spline and wavelet sieve spaces that permits this sharp bound is their local support (see the appendix to Huang (2003)). Other sieve bases such as orthogonal polynomial bases do not have this property and are therefore unable to attain the optimal sup-norm convergence rates for NPIV or nonparametric series LS regression.

Theorem 3.1 *Let Assumptions 1, 2, 3 (with $\Psi_J = \text{BSpl}(J, [0, 1]^d, \gamma)$ or $\text{Wav}(J, [0, 1]^d, \gamma)$), 4 (with $B_K = \text{BSpl}(K, [0, 1]^d, \gamma_x)$ or $\text{Wav}(K, [0, 1]^d, \gamma_x)$), 5 and 7 hold. If $\{(X_i, Y_{2i})\}_{i=1}^n$ is i.i.d. then:*

$$\|h_0 - \hat{h}\|_\infty = O_p(J^{-p/d} + \sigma_{JK}^{-1} \sqrt{K(\log n)/n})$$

provided $J \leq K$, $K \lesssim (n/\log n)^{\delta/(2+\delta)}$, and $\sigma_{JK}^{-1} K \sqrt{(\log n)/n} \lesssim 1$ as $n, J, K \rightarrow \infty$.

(1) *Mildly ill-posed case ($\sigma_{JK}^{-1} = O(J^{\varsigma/d})$ or $\varphi(t) = t^\varsigma$). If Assumption 2 holds with $\delta \geq d/(\varsigma+p)$, and $J \asymp K \asymp (n/\log n)^{d/(2(p+\varsigma)+d)}$ with $K/J \rightarrow c_0 \geq 1$, then:*

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

(2) *Severely ill-posed case ($\sigma_{JK}^{-1} = O(\exp(\frac{1}{2}J^{\varsigma/d}))$ or $\varphi(t) = \exp(-t^{-\varsigma/2})$). If Assumption 2 holds with $\delta > 0$, and $J = c'_0(\log n)^{d/\varsigma}$ for any $c'_0 \in (0, 1)$ with $K = c_0 J$ for some finite $c_0 \geq 1$, then:*

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

Remark 3.2 *Under conditions similar to those for Theorem 3.1, Blundell, Chen, and Kristensen (2007), Chen and Reiss (2011) and Chen and Pouzo (2012) previously obtained the following $L^2(Y_2)$ -norm convergence rate for the sieve NPIV estimator:*

$$\|h_0 - \hat{h}\|_{L^2(Y_2)} = O_p(J^{-p/d} + \tau_{2,2,J} \sqrt{K/n}).$$

(1) *Mildly ill-posed case ($\tau_{2,2,J} = O(J^{\varsigma/d})$ or $\varphi(t) = t^\varsigma$),*

$$\|h_0 - \hat{h}\|_{L^2(Y_2)} = O_p(n^{-p/(2(p+\varsigma)+d)}).$$

(2) *Severely ill-posed case ($\tau_{2,2,J} = O(\exp(\frac{1}{2}J^{\varsigma/d}))$ or $\varphi(t) = \exp(-t^{-\varsigma/2})$),*

$$\|h_0 - \hat{h}\|_{L^2(Y_2)} = O_p((\log n)^{-p/\varsigma}).$$

Chen and Reiss (2011) show that these $L^2(Y_2)$ -norm rates are optimal in the sense that they coincide with the minimax risk lower bound in $L^2(Y_2)$ loss. It is interesting to see that our sup-norm convergence rate is the same as the known optimal $L^2(Y_2)$ -norm rate for the severely ill-posed case, and is only power of $\log(n)$ slower than the known optimal $L^2(Y_2)$ -norm rate for the mildly ill-posed case. In the next subsection we will show that our sup-norm convergence rates are in fact optimal as well.

3.2 Lower bounds on uniform convergence rates for NPIR and NPIV models

For severely ill-posed NPIV models, Chen and Reiss (2011) already showed that $(\log n)^{-p/\varsigma}$ is the minimax lower bound in $L^2(Y_2)$ -norm loss uniformly over a class of functions that include

the Hölder ball $B(p, L)$ as a subset. Therefore, we have for a severely ill-posed NPIV model with $\delta_n = (\log n)^{-p/\varsigma}$,

$$\inf_{\tilde{h}_n} \sup_{h \in B(p, L)} \mathbb{P}_h \left(\|h - \tilde{h}_n\|_\infty \geq c\delta_n \right) \geq \inf_{\tilde{h}_n} \sup_{h \in B(p, L)} \mathbb{P}_h \left(\|h - \tilde{h}_n\|_{L^2(Y_2)} \geq c\delta_n \right) \geq c'$$

where $\inf_{\tilde{h}_n}$ denotes the infimum over all estimators based on a random sample of size n drawn from the NPIV model, and the finite positive constants c, c' do not depend on sample size n . This and Remark 3.2(2) together imply that the sieve NPIV estimator attains the optimal uniform convergence rate in the severely ill-posed case.

We next show that the sup-norm rate for the sieve NPIV estimator obtained in the mildly ill-posed case is also optimal. We begin by placing a primitive smoothness condition on the conditional expectation operator $T : L^2(Y_2) \rightarrow L^2(X)$.

Assumption 8 *There is a $\varsigma > 0$ such that $\|Th\|_{L^2(X)} \lesssim \|h\|_{B_{2,2}^{-\varsigma}}$ for all $h \in B(p, L)$.*

Assumption 8 is a special case of the so-called “link condition” in Chen and Reiss (2011) for the mildly ill-posed case. It can be equivalently stated as: $\|Th\|_{L^2(X)}^2 \lesssim \sum_{j=1}^{\infty} \varphi(j^{-2/d}) |E[h(Y_2)\tilde{\psi}_{J_j}(Y_2)]|^2$ for all $h \in B(p, L)$, with $\varphi(t) = t^\varsigma$ for the mildly ill-posed case. Under this assumption, $n^{-p/(2(p+\varsigma)+d)}$ is the minimax risk lower bound uniformly over the Hölder ball $B(p, L)$ in $L^2(Y_2)$ -norm loss for the mildly ill-posed NPIR and NPIV models (see Chen and Reiss (2011)). We next establish the corresponding minimax risk lower bound in sup-norm loss.

Theorem 3.2 *Let Assumption 8 hold for the NPIV model with a random sample $\{(Y_{1i}, Y_{2i}, X_i)\}_{i=1}^n$. Then:*

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

where $\inf_{\tilde{h}_n}$ denotes the infimum over all estimators based on the sample of size n , and the finite positive constants c, c' do not depend on n .

As in Chen and Reiss (2011), Theorem 3.2 is proved by (i) noting that the risk (in sup-norm loss) for the NPIV model is at least as large as the risk (in sup-norm loss) for the NPIR model, and (ii) calculating a lower bound (in sup-norm loss) for the NPIR model. We consider a Gaussian reduced-form NPIR model with known operator T , given by

$$\begin{aligned} Y_{1i} &= Th_0(X_i) + u_i, \quad i = 1, \dots, n, \\ u_i | X_i &\sim N(0, \sigma^2(X_i)) \quad \text{with} \quad \inf_x \sigma^2(x) \geq \sigma_0^2 > 0. \end{aligned} \tag{20}$$

Theorem 3.2 therefore follows from a sup-norm analogue of Lemma 1 of Chen and Reiss (2011) and the following theorem, which establishes a lower bound on minimax risk over Hölder classes under sup-norm loss for the NPIR model.

Theorem 3.3 *Let Assumption 8 hold for the NPIR model (20) with a random sample $\{(Y_{1i}, X_i)\}_{i=1}^n$. Then:*

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

where $\inf_{\tilde{h}_n}$ denotes the infimum over all estimators based on the sample of size n , and the finite positive constants c, c' depend only on p, L, d, ς and σ_0 .

4 Optimal uniform convergence rates for sieve LS estimators

The standard nonparametric regression model can be recovered as a special case of (4) in which there is no endogeneity, i.e. $Y_2 = X$ and

$$\begin{aligned} Y_{1i} &= h_0(X_i) + \epsilon_i \\ E[\epsilon_i | X_i] &= 0 \end{aligned} \tag{21}$$

in which case $h_0(x) = E[Y_{1i} | X_i = x]$.

Stone (1982) (also see Tsybakov (2009)) establishes that $(n/\log n)^{-p/(2p+d)}$ is the minimax risk lower bound in sup-norm loss for the nonparametric LS regression model (21) with $h_0 \in B(p, L)$. In this section we apply the general upper bound (Theorem 2.1) to show that spline and wavelet sieve LS estimators attain this minimax lower bound for weakly dependent data allowing for heavy-tailed error terms ϵ_i .

Our proof proceeds by noticing that the sieve LS regression estimator

$$\hat{h}(x) = b^K(x)(B'B)^{-1}B'Y \tag{22}$$

obtains as a special case of the NPIV estimator by setting $Y_2 = X$, $\psi^J = b^K$, $J = K$ and $\gamma = \gamma_x$. In this setting, the quantity $P_n h_0(x)$ just reduces to the orthogonal projection of h_0 onto the sieve space B_K under the inner product induced by the empirical distribution, viz.

$$P_n h_0(x) = \tilde{b}^K(x)(\tilde{B}'\tilde{B}/n)^{-1}\tilde{B}'H_0/n. \tag{23}$$

Moreover, in this case the $J \times K$ matrix S defined in (12) reduces to the $K \times K$ identity matrix I_K and its smallest singular value is unity (whence $\sigma_{JK} = 1$). Therefore, the general calculation presented in Theorem 2.1 can be used to control the ‘‘variance term’’ $\|\hat{h} - P_n h_0\|_\infty$. The ‘‘bias term’’ $\|P_n h_0 - h_0\|_\infty$ is controlled as in Huang (2003). It is worth emphasizing that no explicit weak dependence condition is placed on the regressors $\{X_i\}_{i=-\infty}^\infty$. Instead, this is implicitly captured by Condition (ii) on convergence of $\tilde{B}'\tilde{B}/n - I_K$.

Theorem 4.1 *Let Assumptions 1, 2, 4 (with $B_K = BSpl(K, [0, 1]^d, \gamma)$ or $Wav(K, [0, 1]^d, \gamma)$) and 5 hold for Model (21). Then:*

$$\|\widehat{h} - h_0\|_\infty = O_p(K^{-p/d} + \sqrt{K(\log n)/n})$$

provided $n, K \rightarrow \infty$, and

(i) $K \lesssim (n/\log n)^{\delta/(2+\delta)}$ and $\sqrt{K(\log n)/n} \lesssim 1$

(ii) $\|(\widetilde{B}'\widetilde{B}/n) - I_K\| = O_p(\sqrt{(\log n)/K}) = o_p(1)$.

Condition (ii) is satisfied by applying Lemma 5.2 for i.i.d. data and Lemma 5.3 for weakly dependent data. Theorem 4.1 shows that spline and wavelet sieve LS estimators can achieve this minimax lower bound for weakly dependent data.

Corollary 4.1 *Let Assumptions 1, 2 (with $\delta \geq d/p$), 4 (with $B_K = BSpl(K, [0, 1]^d, \gamma)$ or $Wav(K, [0, 1]^d, \gamma)$) and 5 hold for Model (21). If $K \asymp (n/\log n)^{d/(2p+d)}$ then:*

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

provided that one of the followings is satisfied

- (1) the regressors are i.i.d.;
- (2) the regressors are exponentially β -mixing and $d < 2p$;
- (3) the regressors are algebraically β -mixing at rate γ and $(2 + \gamma)d < 2\gamma p$.

Corollary 4.1 states that for i.i.d. data, Stone's optimal sup-norm convergence rate is achieved by spline and wavelet LS estimators whenever $\delta \geq d/p$ and $d \leq 2p$ (Assumption 5). If the regressors are exponentially β -mixing the optimal rate of convergence is achieved with $\delta \geq d/p$ and $d < 2p$. The restrictions $\delta \geq d/p$ and $(2 + \gamma)d < 2\gamma p$ for algebraically β -mixing (at a rate γ) reduces naturally towards the exponentially mixing conditions as the dependence becomes weaker (i.e. γ becomes larger). In all cases, a smoother function (i.e., bigger p) means a lower value of δ , and therefore heavier-tailed error terms ϵ_i , are permitted while still obtaining the optimal sup-norm convergence rate. In particular this is achieved with $\delta = d/p \leq 2$ for i.i.d. data. Recently, Belloni et al. (2012) require that the conditional $(2 + \eta)$ th moment (for some $\eta > 0$) of ϵ_i be uniformly bounded for spline LS regression estimators to achieve the optimal sup-norm rate for i.i.d. data.⁵

⁵Chen would like to thank Jianhua Huang for working together on an earlier draft that does achieve the optimal sup-norm rate for a polynomial spline LS estimator with i.i.d. data, but under a stronger condition that $E[\epsilon_i^4 | X_i = x]$ is uniformly bounded in x .

Uniform convergence rates of series LS estimators have also been studied by Newey (1997), de Jong (2002), Song (2008), Lee and Robinson (2013) and others, but the sup-norm rates obtained in these papers are slower than the minimax risk lower bound in sup-norm loss of Stone (1982).⁶ Our result is the first such optimal sup-norm rate result for a sieve nonparametric LS estimator allowing for weakly-dependent data with heavy-tailed error terms. It should be very useful for nonparametric estimation of financial time-series models that have heavy-tailed error terms.

5 Useful results on random matrices

5.1 Convergence rates for sums of dependent random matrices

In this subsection a Bernstein inequality for sums of independent random matrices due to Tropp (2012) is adapted to obtain convergence rates for sums of random matrices formed from β -mixing (absolutely regular) sequences, where the dimension, norm, and variance measure of the random matrices are allowed to grow with the sample size. These inequalities are particularly useful for establishing convergence rates for semi/nonparametric sieve estimators with weakly-dependent data. We first recall a result of Tropp (2012).

Theorem 5.1 (Tropp (2012)) *Let $\{\Xi_i\}_{i=1}^n$ be a finite sequence of independent random matrices with dimensions $d_1 \times d_2$. Assume $E[\Xi_i] = 0$ for each i and $\max_{1 \leq i \leq n} \|\Xi_i\| \leq R_n$, and define*

$$\sigma_n^2 = \max \left\{ \left\| \sum_{i=1}^n E[\Xi_i \Xi_i'] \right\|, \left\| \sum_{i=1}^n E[\Xi_i' \Xi_i] \right\| \right\}.$$

Then for all $t \geq 0$,

$$\mathbb{P} \left(\left\| \sum_{i=1}^n \Xi_i \right\| \geq t \right) \leq (d_1 + d_2) \exp \left(\frac{-t^2/2}{\sigma_n^2 + R_n t/3} \right).$$

Corollary 5.1 *Under the conditions of Theorem 5.1, if $R_n \sqrt{\log(d_1 + d_2)} = o(\sigma_n)$ then*

$$\left\| \sum_{i=1}^n \Xi_{i,n} \right\| = O_p(\sigma_n \sqrt{\log(d_1 + d_2)}).$$

We now provide a version of Theorem 5.1 and Corollary 5.1 for matrix-valued functions of β -mixing sequences. The β -mixing coefficient between two σ -algebras \mathcal{A} and \mathcal{B} is defined as

$$2\beta(\mathcal{A}, \mathcal{B}) = \sup_{(i,j) \in I \times J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)| \quad (24)$$

⁶See, e.g., Hansen (2008), Masry (1996), Cattaneo and Farrell (2013) and the references therein for the optimal sup-norm convergence rates of a conditional mean function via the kernel, local linear regression and partitioning estimators of a conditional mean function.

with the supremum taken over all finite partitions $\{A_i\}_{i \in I} \subset \mathcal{A}$ and $\{B_j\}_{j \in J} \subset \mathcal{B}$ (Doukhan, Massart, and Rio, 1995). The q th β -mixing coefficient of $\{X_i\}_{i=-\infty}^{\infty}$ is defined as

$$\beta(q) = \sup_i \beta(\sigma(\dots, X_{i-1}, X_i), \sigma(X_{i+q}, X_{i+q+1}, \dots)). \quad (25)$$

The process $\{X_i\}_{i=-\infty}^{\infty}$ is said to be *algebraically β -mixing* at rate γ if $q^\gamma \beta(q) = o(1)$ for some $\gamma > 1$, and *geometrically β -mixing* if $\beta(q) \leq c \exp(-\gamma q)$ for some $\gamma > 0$ and $c \geq 0$. The following extension of Theorem 5.1 is made using a Berbee's lemma and a coupling argument (see, e.g., Doukhan et al. (1995)).

Theorem 5.2 *Let $\{X_i\}_{i=-\infty}^{\infty}$ be a strictly stationary β -mixing sequence and let $\Xi_{i,n} = \Xi_n(X_i)$ for each i where $\Xi_n : \mathcal{X} \rightarrow \mathbb{R}^{d_1 \times d_2}$ is a sequence of measurable $d_1 \times d_2$ matrix-valued functions. Assume $E[\Xi_{i,n}] = 0$ and $\|\Xi_{i,n}\| \leq R_n$ for each i and define $s_n^2 = \max_{1 \leq i, j \leq n} \max\{\|E[\Xi_{i,n} \Xi'_{j,n}]\|, \|E[\Xi'_{i,n} \Xi_{j,n}]\|\}$. Let q be an integer between 1 and $n/2$ and let $I_r = q[n/q] + 1, \dots, n$ when $q[n/q] < n$ and $I_r = \emptyset$ when $q[n/q] = n$. Then for all $t \geq 0$,*

$$\mathbb{P} \left(\left\| \sum_{i=1}^n \Xi_{i,n} \right\| \geq 6t \right) \leq \frac{n}{q} \beta(q) + \mathbb{P} \left(\left\| \sum_{i \in I_r} \Xi_{i,n} \right\| \geq t \right) + 2(d_1 + d_2) \exp \left(\frac{-t^2/2}{nq s_n^2 + qR_n t/3} \right)$$

(where $\|\sum_{i \in I_r} \Xi_{i,n}\| := 0$ whenever $I_r = \emptyset$).

Corollary 5.2 *Under the conditions of Theorem 5.2, if $q = q(n)$ is chosen such that $\frac{n}{q} \beta(q) = o(1)$ and $R_n \sqrt{q \log(d_1 + d_2)} = o(s_n \sqrt{n})$ then*

$$\left\| \sum_{i=1}^n \Xi_{i,n} \right\| = O_p(s_n \sqrt{nq \log(d_1 + d_2)}).$$

5.2 Empirical identifiability

This subsection provides a readily verifiable condition under which, with probability approaching one (wpa1), the theoretical and empirical L^2 norms are equivalent over a linear sieve space. This equivalence, referred to by Huang (2003) as *empirical identifiability*, has several applications in nonparametric sieve estimation. In the context of nonparametric series regression, empirical identifiability ensures the estimator is the orthogonal projection of Y onto the sieve space under the empirical inner product and is uniquely defined (Huang, 2003). Empirical identifiability is also used to establish the large-sample properties of sieve conditional moment estimators (Chen and Pouzo, 2012). A sufficient condition for empirical identifiability is now cast in terms of convergence of a random matrix, which we verify for i.i.d. and β -mixing sequences.

A subspace $\mathcal{A} \subseteq L^2(X)$ is said to be *empirically identifiable* if $\frac{1}{n} \sum_{i=1}^n b(X_i)^2 = 0$ implies $b = 0$ a.e.- $[F_X]$ where F_X denotes the distribution of X . A sequence of spaces $\{\mathcal{A}_K : K \geq 1\} \subseteq L^2(X)$ is

empirically identifiable wpa1 as $K = K(n) \rightarrow \infty$ with n if

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\sup_{a \in A_K} \left| \frac{\frac{1}{n} \sum_{i=1}^n a(X_i)^2 - E[a(X)^2]}{E[a(X)^2]} \right| > t \right) = 0 \quad (26)$$

for any $t > 0$. Huang (1998) uses a chaining argument to provide sufficient conditions for (26) over the linear space B_K under i.i.d. sampling. Chen and Pouzo (2012) use this argument to establish convergence of sieve conditional moment estimators. Although easy to establish for i.i.d. sequences, it may be difficult to verify (26) via chaining arguments for certain types of weakly dependent sequences. To this end, the following is a readily verifiable sufficient condition for empirical identifiability for linear sieve spaces. Let $B_K = \text{clsp}\{b_{K1}, \dots, b_{KK}\}$ denote a general linear sieve space and let $\tilde{B} = (\tilde{b}^K(X_1), \dots, \tilde{b}^K(X_n))'$ where $\tilde{b}^K(x)$ is the orthonormalized vector of basis functions.

Condition 5.1 $\lambda_{\min}(E[b^K(X)b^K(X)']) > 0$ for each $K \geq 1$ and $\|\tilde{B}'\tilde{B}/n - I_K\| = o_p(1)$.

Lemma 5.1 If $\lambda_{\min}(E[b^K(X)b^K(X)']) > 0$ for each $K \geq 1$ then

$$\sup_{b \in B_K} \left| \frac{\frac{1}{n} \sum_{i=1}^n b(X_i)^2 - E[b(X)^2]}{E[b(X)^2]} \right| = \|\tilde{B}'\tilde{B}/n - I_K\|^2.$$

Corollary 5.3 Under Condition 5.1, B_K is empirically identifiable wpa1.

Condition 5.1 is a sufficient condition for (26) with a linear sieve space B_K . It should be noted that convergence is only required in the spectral norm. In the i.i.d. case this allows for K to increase more quickly with n than is achievable under the chaining argument of Huang (1998). Let

$$\zeta_0(K) = \sup_{x \in \mathcal{X}} \|b^K(x)\| \quad (27)$$

as in Newey (1997). Under regularity conditions, $\zeta_0(K) = O(\sqrt{K})$ for tensor products of splines, trigonometric polynomials or wavelets and $\zeta_0(K) = O(K)$ for tensor products of power series or polynomials (Newey, 1997; Huang, 1998). Under the chaining argument of Huang (1998), (26) is achieved under the restriction $\zeta_0(K)^2 K/n = o(1)$. Huang (2003) relaxes this restriction to $K(\log n)/n = o(1)$ for a polynomial spline sieve. We now generalize this result by virtue of Lemma 5.1 and exponential inequalities for sums of random matrices.

Lemma 5.2 If $\{X_i\}_{i=1}^n$ is i.i.d. and $\lambda_{\min}(E[b^K(X)b^K(X)']) \geq \underline{\lambda} > 0$ for each $K \geq 1$, then

$$\|(\tilde{B}'\tilde{B}/n) - I_K\| = O_p(\zeta_0(K)\sqrt{(\log K)/n})$$

provided $\zeta_0(K)^2(\log K)/n = o(1)$.

Remark 5.1 If $\{X_i\}_{i=1}^n$ is i.i.d., $K(\log K)/n = o(1)$ is sufficient for sieve bases that are tensor products of splines, trigonometric polynomials or wavelets, and $K^2(\log K)/n = o(1)$ is sufficient for sieve bases that are tensor products of power series or polynomials.

The following lemma is useful to provide sufficient conditions for empirical identifiability for β -mixing sequences, which uses Theorem 5.2.

Lemma 5.3 If $\{X_i\}_{i=-\infty}^{\infty}$ is strictly stationary and β -mixing with mixing coefficients such that one can choose an integer sequence $q = q(n) \leq n/2$ with $\beta(q)n/q = o(1)$ and $\lambda_{\min}(E[b^K(X)b^K(X)']) \geq \lambda > 0$ for each $K \geq 1$, then

$$\|(\tilde{B}'\tilde{B}/n) - I_K\| = O_p(\zeta_0(K)\sqrt{q(\log K)/n})$$

provided $\zeta_0(K)^2q \log K/n = o(1)$.

Remark 5.2 If $\{X_i\}_{i=-\infty}^{\infty}$ is algebraically β -mixing at rate γ , $Kn^{1/(1+\gamma)}(\log K)/n = o(1)$ is sufficient for sieve bases that are tensor products of splines, trigonometric polynomials or wavelets, and $K^2n^{1/(1+\gamma)}(\log K)/n = o(1)$ is sufficient for sieve bases that are tensor products of power series or polynomials.

Remark 5.3 If $\{X_i\}_{i=-\infty}^{\infty}$ is geometrically β -mixing, $K(\log n)^2/n = o(1)$ is sufficient for sieve bases that are tensor products of splines, trigonometric polynomials or wavelets, and $K^2(\log n)^2/n = o(1)$ is sufficient for sieve bases that are tensor products of power series or polynomials.

A Brief review of B-spline and wavelet sieve spaces

We first outline univariate B-spline and wavelet sieve spaces on $[0, 1]$, then deal with the multivariate case by constructing a tensor-product sieve basis.

B-splines B-splines are defined by their order $m \geq 1$ and number of interior knots $N \geq 0$. Define the knot set

$$t_{-(m-1)} = \dots = t_0 \leq t_1 \leq \dots \leq t_N \leq t_{N+1} = \dots = t_{N+m} \quad (28)$$

where we normalize $t_0 = 0$ and $t_{N+1} = 1$. The B-spline basis is then defined recursively via the De Boor relation. This results in a total of $K = N + m$ splines which together form a partition of unity. Each spline is a polynomial of degree $m - 1$ on each interior interval $I_1 = [t_0, t_1), \dots, I_n = [t_N, t_{N+1}]$ and is $(m - 2)$ -times continuously differentiable on $[0, 1]$ whenever $m \geq 2$. The mesh ratio is defined as

$$\text{mesh}(K) = \frac{\max_{0 \leq n \leq N} (t_{n+1} - t_n)}{\min_{0 \leq n \leq N} (t_{n+1} - t_n)}. \quad (29)$$

We let the space $\text{BSpl}(K, [0, 1])$ be the closed linear span of these $K = N + m$ splines. The space $\text{BSpl}(K, [0, 1])$ has *uniformly bounded mesh ratio* if $\text{mesh}(K) \leq \kappa$ for all $N \geq 0$ and some $\kappa \in (0, \infty)$. The space $\text{BSpl}(K, [0, 1])$ has *smoothness* $\gamma = m - 2$, which is denoted as $\text{BSpl}(K, [0, 1], \gamma)$ for simplicity. See De Boor (2001) and Schumacker (2007) for further details.

Wavelets We follow the construction of Cohen et al. (1993a,b) for building a wavelet basis for $[0, 1]$. Let (ϕ, ψ) be a father and mother wavelet pair that has N vanishing moments and $\text{support}(\phi) = \text{support}(\psi) = [0, 2N - 1]$. For given j , the approximation space V_j wavelet space W_j each consist of 2^j functions $\{\phi_{jk}\}_{1 \leq k \leq 2^j}$ and $\{\psi_{jk}\}_{1 \leq k \leq 2^j}$ respectively, such that $\{\phi_{jk}\}_{1 \leq k \leq 2^{j-2N}}$ and $\{\psi_{jk}\}_{1 \leq k \leq 2^{j-2N}}$ are interior wavelets for which $\phi_{jk}(\cdot) = 2^{j/2}\phi(2^j(\cdot) - k)$ and $\psi_{jk}(\cdot) = 2^{j/2}\psi(2^j(\cdot) - k)$, complemented with another N left-edge functions and N right-edge functions. Choosing $L \geq 1$ such that $2^L \geq 2N$, we let the space $\text{Wav}(K, [0, 1])$ be the closed linear span of the set of functions

$$W_{LJ} = \{\phi_{Lk} : 1 \leq k \leq 2^L\} \cup \{\psi_{jk} : k = 1, \dots, 2^j \text{ and } j = L, \dots, J - 1\} \quad (30)$$

for integer $J > L$, and let $K = \#(W_{LJ})$. We say that $\text{Wav}(K, [0, 1])$ has *regularity* γ if ϕ and ψ are both γ times continuously differentiable, which is denoted as $\text{Wav}(K, [0, 1], \gamma)$ for simplicity.

Tensor products To construct a tensor-product B-spline basis of smoothness γ for $[0, 1]^d$ with $d > 1$, we first construct d univariate B-spline bases for $[0, 1]$, say G_i with $G_i = \text{BSpl}(k, [0, 1])$ and smoothness γ for each $1 \leq i \leq d$. We then set $K = k^d$ and let $\text{BSpl}(K, [0, 1]^d)$ be spanned by the unique k^d functions given by $\prod_{i=1}^d g_i$ with $g_i \in G_i$ for $1 \leq i \leq d$. The tensor-product wavelet basis $\text{Wav}(K, [0, 1]^d)$ of regularity γ for $[0, 1]^d$ is formed similarly as the tensor product of d univariate Wavelet bases of regularity γ (see Triebel (2006, 2008)).

Wavelet characterization of Besov norms Let $f \in B_{p,q}^\alpha([0, 1]^d)$ have wavelet expansion

$$f = \sum_{k=-\infty}^{\infty} \mathbf{a}_k(f) \phi_{Lk} + \sum_{j=L}^{\infty} \sum_{k=-\infty}^{\infty} \mathbf{b}_{jk}(f) \psi_{jk} \quad (31)$$

where $\{\phi_{Lk}, \psi_{jk}\}_{j,k}$ are a Wavelet basis with regularity $\gamma > \alpha$. Equivalent norms to the $B_{\infty,\infty}^\alpha$ and $B_{2,2}^\alpha$ norms may be formulated equivalently in terms of the wavelet coefficient sequences $\{\mathbf{a}_k\}_k^\infty$ and $\{\mathbf{b}_{jk}\}_{j,k}$, namely $\|\cdot\|_{b_{\infty,\infty}^\alpha}$ and $\|\cdot\|_{b_{2,2}^\alpha}$, given by

$$\begin{aligned} \|f\|_{b_{\infty,\infty}^\alpha} &= \sup_k |\mathbf{a}_k(f)| + \sup_{j,k} 2^{j(\alpha+d/2)} |\mathbf{b}_{jk}(f)| \\ \|f\|_{b_{2,2}^\alpha} &= \|\mathbf{a}_{(\cdot)}(f)\| + \left(\sum_{j=0}^{\infty} (2^{j\alpha} \|\mathbf{b}_{j(\cdot)}(f)\|)^2 \right)^{1/2} \end{aligned} \quad (32)$$

where $\|\mathbf{a}_{(\cdot)}(f)\|$ and $\|\mathbf{b}_{j(\cdot)}(f)\|$ denote the infinite-dimensional Euclidean norm for the sequences $\{\mathbf{a}_k(f)\}_k$ and $\{\mathbf{b}_{jk}(f)\}_k$ (see, e.g., Johnstone (2013) and Triebel (2006, 2008)).

B Proofs of main results

B.1 Proofs for Section 2

Proof of Theorem 2.1. It is enough to show that $\|\hat{h} - P_n h_0\|_\infty = O_p(\sigma_{JK}^{-1} \sqrt{K(\log n)/n})$. First write

$$\hat{h}(y_2) - P_n h_0(y_2) = \tilde{\psi}^J(y_2)' [\hat{S}(\tilde{B}' \tilde{B}/n) - \hat{S}'] - \hat{S}(\tilde{B}' \tilde{B}/n) - \tilde{B}' e/n \quad (33)$$

where $e = (\epsilon_1, \dots, \epsilon_n)$. Convexity of \mathcal{Y}_2 (Assumption 1(ii)), smoothness of $\tilde{\psi}^J$ (Assumption 3(i)) and the mean value theorem provide that, for any $(y, y^*) \in \mathcal{Y}_2^2$,

$$\begin{aligned} |\hat{h}(y) - P_n h_0(y) - (\hat{h}(y^*) - P_n h_0(y^*))| &= |(\tilde{\psi}^J(y) - \tilde{\psi}^J(y^*))' [\hat{S}(\tilde{B}'\tilde{B}/n)^- \hat{S}'^- \hat{S}(\tilde{B}'\tilde{B}/n)^- \tilde{B}'e/n]| \quad (34) \\ &= |(y - y^*)' \nabla \tilde{\psi}^J(y^{**})' [\hat{S}(\tilde{B}'\tilde{B}/n)^- \hat{S}'^- \hat{S}(\tilde{B}'\tilde{B}/n)^- \tilde{B}'e/n]| \quad (35) \\ &\leq J^\alpha \|y - y^*\| \|[\hat{S}(\tilde{B}'\tilde{B}/n)^- \hat{S}'^- \hat{S}(\tilde{B}'\tilde{B}/n)^- \tilde{B}'e/n]\| \quad (36) \end{aligned}$$

for some y^{**} in the segment between y and y^* , and some $\alpha > 0$ (and independent of y and y^*).

We first show that $T_1 := \|[\hat{S}(\tilde{B}'\tilde{B}/n)^- \hat{S}'^- \hat{S}(\tilde{B}'\tilde{B}/n)^- \tilde{B}'e/n]\| = o_p(1)$. By the triangle inequality and properties of the matrix spectral norm,

$$T_1 \leq (\|[\hat{S}(\tilde{B}'\tilde{B}/n)^- \hat{S}'^- \hat{S}(\tilde{B}'\tilde{B}/n)^- - [SS']^{-1}S]\| + \|[SS']^{-1}S\|) \|\tilde{B}'e/n\| \quad (37)$$

whence by Lemma C.2 (under condition (ii) of the Theorem), wpa1

$$T_1 \lesssim \left\{ \sigma_{JK}^{-1} \|(\tilde{B}'\tilde{B}/n) - I_K\| + \sigma_{JK}^{-2} \left(\|\hat{S} - S\| + \|(\tilde{B}'\tilde{B}/n) - I_K\| \right) + \sigma_{JK}^{-1} \right\} \|\tilde{B}'e/n\|. \quad (38)$$

Noting that $\|\tilde{B}'e/n\| = O_p(\sqrt{K/n})$ (by Markov's inequality under Assumptions 2 and 4), it follows by conditions (i) and (ii) of the Theorem that $T_1 = o_p(1)$. Therefore, for any fixed $\bar{M} > 0$ we have

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left(\|[\hat{S}(\tilde{B}'\tilde{B}/n)^- \hat{S}'^- \hat{S}(\tilde{B}'\tilde{B}/n)^- \tilde{B}'e/n]\| > \bar{M} \right) = 0. \quad (39)$$

Let \mathcal{B}_n denote the event $\|[\hat{S}(\tilde{B}'\tilde{B}/n)^- \hat{S}'^- \hat{S}(\tilde{B}'\tilde{B}/n)^- \tilde{B}'e/n]\| \leq \bar{M}$ and observe that $\mathbb{P}(\mathcal{B}_n^c) = o(1)$. On \mathcal{B}_n , for any $C \geq 1$, a finite positive $\beta = \beta(C)$ and $\gamma = \gamma(C)$ can be chosen such that

$$J^\alpha \|y_0 - y_1\| \|[\hat{S}(\tilde{B}'\tilde{B}/n)^- \hat{S}'^- \hat{S}(\tilde{B}'\tilde{B}/n)^- \tilde{B}'e/n]\| \leq C \sigma_{JK}^{-1} \sqrt{K(\log n)/n} \quad (40)$$

whenever $\|y_0 - y_1\| \leq \beta n^{-\gamma}$. Let \mathcal{S}_n be the smallest subset of \mathcal{Y}_2 such that for each $y \in \mathcal{Y}_2$ there exists a $y_n \in \mathcal{S}_n$ with $\|y_n - y\| \leq \beta n^{-\gamma}$. For any $y \in \mathcal{Y}_2$ let $y_n(y)$ denote the $y_n \in \mathcal{S}_n$ nearest (in Euclidean distance) to y . Therefore,

$$|\hat{h}(y) - P_n h_0(y) - (\hat{h}(y_n(y)) - P_n h_0(y_n(y)))| \leq C \sigma_{JK}^{-1} \sqrt{K(\log n)/n} \quad (41)$$

for any $y \in \mathcal{Y}_2$, on \mathcal{B}_n .

For any $C \geq 1$, straightforward arguments yield

$$\begin{aligned} &\mathbb{P} \left(\|\hat{h} - P_n h_0\|_\infty \geq 4C \sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right) \\ &\leq \mathbb{P} \left(\left\{ \|\hat{h} - P_n h_0\|_\infty \geq 4C \sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right\} \cap \mathcal{B}_n \right) + \mathbb{P}(\mathcal{B}_n^c) \quad (42) \end{aligned}$$

$$\begin{aligned} &\leq \mathbb{P} \left(\left\{ \sup_{y \in \mathcal{Y}_2} |\hat{h}(y) - P_n h_0(y) - (\hat{h}(y_n(y)) - P_n h_0(y_n(y)))| \geq 2C \sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right\} \cap \mathcal{B}_n \right) \\ &\quad + \mathbb{P} \left(\left\{ \max_{y_n \in \mathcal{S}_n} |\hat{h}(y_n) - P_n h_0(y_n)| \geq 2C \sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right\} \cap \mathcal{B}_n \right) + \mathbb{P}(\mathcal{B}_n^c) \quad (43) \end{aligned}$$

$$= \mathbb{P} \left(\left\{ \max_{y_n \in \mathcal{S}_n} |\hat{h}(y_n) - P_n h_0(y_n)| \geq 2C \sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right\} \cap \mathcal{B}_n \right) + o(1) \quad (44)$$

where the final line is by (41) and the fact that $\mathbb{P}(\mathcal{B}_n^c) = o(1)$. For the remaining term:

$$\begin{aligned} & \mathbb{P} \left(\left\{ \max_{y_n \in \mathcal{S}_n} |\widehat{h}(y_n) - P_n h_0(y_n)| \geq 2C\sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right\} \cap \mathcal{B}_n \right) \\ & \leq \mathbb{P} \left(\max_{y_n \in \mathcal{S}_n} |\widetilde{\psi}^J(y_n)' [\widehat{S}(\widetilde{B}'\widetilde{B}/n)^- \widehat{S}'^- \widehat{S}(\widetilde{B}'\widetilde{B}/n)^- \widetilde{B}'e/n] \geq 2C\sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right) \end{aligned} \quad (45)$$

$$\leq \mathbb{P} \left(\max_{y_n \in \mathcal{S}_n} |\widetilde{\psi}^J(y_n)' \{[\widehat{S}(\widetilde{B}'\widetilde{B}/n)^- \widehat{S}'^- \widehat{S}(\widetilde{B}'\widetilde{B}/n)^- - [SS']^{-1}S\} \widetilde{B}'e/n| \geq C\sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right) \quad (46)$$

$$+ \mathbb{P} \left(\max_{y_n \in \mathcal{S}_n} |\widetilde{\psi}^J(y_n)' [SS']^{-1}S \widetilde{B}'e/n| \geq C\sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right). \quad (47)$$

It is now shown that a sufficiently large C can be chosen to make terms (46) and (47) arbitrarily small as $n, J, K \rightarrow \infty$. Observe that \mathcal{S}_n has cardinality $\lesssim n^\nu$ for some $\nu = \nu(C) \in (0, \infty)$ under Assumption 1(ii).

Control of (46): The Cauchy-Schwarz inequality and Assumption 3 yield

$$\begin{aligned} & |\widetilde{\psi}^J(y_n)' \{[\widehat{S}(\widetilde{B}'\widetilde{B}/n)^- \widehat{S}'^- \widehat{S}(\widetilde{B}'\widetilde{B}/n)^- - [SS']^{-1}S\} \widetilde{B}'e/n| \\ & \lesssim \sqrt{J} \|[\widehat{S}(\widetilde{B}'\widetilde{B}/n)^- \widehat{S}'^- \widehat{S}(\widetilde{B}'\widetilde{B}/n)^- - [SS']^{-1}S\| \times O_p(\sqrt{K/n}) \end{aligned} \quad (48)$$

uniformly for $y_n \in \mathcal{S}_n$ (recalling that $\|\widetilde{B}'e/n\| = O_p(\sqrt{K/n})$ under Assumptions 2 and 4). Therefore, (46) will vanish asymptotically provided

$$T_2 := \sigma_{JK} \sqrt{J} \|[\widehat{S}(\widetilde{B}'\widetilde{B}/n)^- \widehat{S}'^- \widehat{S}(\widetilde{B}'\widetilde{B}/n)^- - [SS']^{-1}S\| / \sqrt{\log n} = o_p(1). \quad (49)$$

Under condition (ii), the bound

$$T_2 \lesssim \sqrt{J} \left\{ \|(\widetilde{B}'\widetilde{B}/n) - I_K\| + \sigma_{JK}^{-1} \left(\|\widehat{S} - S\| + \|(\widetilde{B}'\widetilde{B}/n) - I_K\| \right) \right\} / \sqrt{\log n} \quad (50)$$

holds wpa1 by Lemma C.2, and so $T_2 = o_p(1)$ by virtue of conditions (i) and (ii) of the Theorem.

Control of (47): Let $\{M_n : n \geq 1\}$ be an increasing sequence diverging to $+\infty$ and define

$$\begin{aligned} \epsilon_{1,i,n} &= \epsilon_i \{|\epsilon_i| \leq M_n\} \\ \epsilon_{2,i,n} &= \epsilon_i - \epsilon_{1,i,n} \\ g_{i,n} &= \psi^J(y_n)' [SS']^{-1} S \widetilde{b}^K(X_i). \end{aligned} \quad (51)$$

Simple application of the triangle inequality yields

$$(47) \leq (\#\mathcal{S}_n) \max_{y_n \in \mathcal{S}_n} \mathbb{P} \left(\left\{ \left| \sum_{i=1}^n g_{i,n} (\epsilon_{1,i,n} - E[\epsilon_{1,i,n} | \mathcal{F}_{i-1}]) \right| > \frac{C}{3} \sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right\} \cap \mathcal{A}_n \right) + \mathbb{P}(\mathcal{A}_n^c) \quad (52)$$

$$+ \mathbb{P} \left(\max_{y_n \in \mathcal{S}_n} \left| \frac{1}{n} \sum_{i=1}^n g_{i,n} E[\epsilon_{1,i,n} | \mathcal{F}_{i-1}] \right| \geq \frac{C}{3} \sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right) \quad (53)$$

$$+ \mathbb{P} \left(\max_{y_n \in \mathcal{S}_n} \left| \frac{1}{n} \sum_{i=1}^n g_{i,n} \epsilon_{2,i,n} \right| \geq \frac{C}{3} \sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right) \quad (54)$$

where \mathcal{A}_n is a measurable set to be defined. The following shows that terms (52), (53), and (54) vanish asymptotically provided a sequence $\{M_n : n \geq 1\}$ may be chosen such that $\sqrt{nJ/\log n} = O(M_n^{1+\delta})$ and $M_n = O(\sqrt{n/(J \log n)})$ and $J \leq K$. Choosing $J \leq K$ and setting $M_n^{1+\delta} \asymp \sqrt{nK/\log n}$ trivially satisfies the

condition $\sqrt{nK/\log n} = O(M_n^{1+\delta})$. The condition $M_n = O(\sqrt{n/(K \log n)})$ is satisfied for this choice of M_n provided $K \lesssim (n/\log n)^{\delta/(2+\delta)}$.

Control of (53) and (54): For term (54), first note that

$$|g_{i,n}| \lesssim \sigma_{JK}^{-1} \sqrt{JK} \quad (55)$$

whenever $\sigma_{JK} > 0$ by the Cauchy-Schwarz inequality, and Assumptions 3 and 4. This, together with Markov's inequality and Assumption 2(iii) yields

$$\mathbb{P} \left(\max_{y_n \in \mathcal{S}_n} \left| \frac{1}{n} \sum_{i=1}^n g_{i,n} \epsilon_{2,i,n} \right| \geq \frac{C}{3} \sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right) \lesssim \frac{\sigma_{JK}^{-1} \sqrt{JK} E[|\epsilon_i| \{|\epsilon_i| > M_n\}]}{\sigma_{JK}^{-1} \sqrt{K(\log n)/n}} \quad (56)$$

$$\leq \sqrt{\frac{nJ}{\log n} \frac{E[|\epsilon_i|^{2+\delta} \{|\epsilon_i| > M_n\}]}{M_n^{1+\delta}}} \quad (57)$$

which is $o(1)$ provided $\sqrt{nJ/\log n} = O(M_n^{1+\delta})$. Term (53) is controlled by an identical argument, using the fact that $E[\epsilon_{1,i,n} | \mathcal{F}_{i-1}] = -E[\epsilon_{2,i,n} | \mathcal{F}_{i-1}]$ by Assumption 2(i).

Control of (52): Term (52) is to be controlled using an exponential inequality for martingales due to van de Geer (1995). Let \mathcal{A}_n denote the set on which $\|(\tilde{B}'\tilde{B}/n) - I_K\| \leq \frac{1}{2}$ and observe that $\mathbb{P}(\mathcal{A}_n^c) = o(1)$ under the condition $\|(\tilde{B}'\tilde{B}/n) - I_K\| = o_p(1)$. Under Assumptions 2(ii), 3, and 4, the predictable variation of the summands in (52) may be bounded by

$$\frac{1}{n^2} \sum_{i=1}^n E[(g_{i,n}(\epsilon_{1,i,n} - E[\epsilon_{1,i,n} | \mathcal{F}_{i-1}]))^2 | \mathcal{F}_{i-1}] \lesssim n^{-1} \tilde{\psi}^J(y_n)' [SS']^{-1} S(\tilde{B}'\tilde{B}/n) S' [SS']^{-1} \tilde{\psi}^J(y_n) \quad (58)$$

$$\lesssim \sigma_{JK}^{-2} J/n \quad \text{on } \mathcal{A}_n \quad (59)$$

uniformly for $y_n \in \mathcal{S}_n$. Moreover, under Assumption 4, each summand is bounded uniformly for $y_n \in \mathcal{S}_n$ by

$$|n^{-1} g_{i,n}(\epsilon_{1,i,n} - E[\epsilon_{1,i,n} | \mathcal{F}_{i-1}])| \lesssim \frac{\sigma_{JK}^{-1} \sqrt{JK} M_n}{n}. \quad (60)$$

Lemma 2.1 of van de Geer (1995) then provides that (52) may be bounded by

$$\begin{aligned} & (\#\mathcal{S}_n) \max_{y_n \in \mathcal{S}_n} \mathbb{P} \left(\left\{ \left| \sum_{i=1}^n g_{i,n}(\epsilon_{1,i,n} - E[\epsilon_{1,i,n} | \mathcal{F}_{i-1}]) \right| > \frac{C}{3} \sigma_{JK}^{-1} \sqrt{K(\log n)/n} \right\} \cap \mathcal{A}_n \right) + \mathbb{P}(\mathcal{A}_n^c) \\ & \lesssim n^\nu \exp \left\{ - \frac{C \sigma_{JK}^{-2} K(\log n)/n}{c_1 \sigma_{JK}^{-2} J/n + c_2 n^{-1} \sigma_{JK}^{-2} \sqrt{JK} M_n \sqrt{CK(\log n)/n}} \right\} + o(1) \end{aligned} \quad (61)$$

$$\lesssim \exp \left\{ \log n - \frac{CK(\log n)/n}{c_3 J/n} \right\} + \exp \left\{ \log n - \frac{\sqrt{CK(\log n)/n}}{c_4 K M_n/n} \right\} + o(1) \quad (62)$$

for finite positive constants c_1, \dots, c_4 . Thus (52) is $o(1)$ for large enough C by virtue of the conditions $M_n = O(\sqrt{n/(J \log n)})$ and $J \leq K$. ■

B.2 Proofs for Section 3

Proof of Theorem 3.1. Theorem 2.1 gives $\|\hat{h} - P_n h_0\|_\infty = O_p(\sigma_{JK}^{-1} \sqrt{K(\log n)/n})$ provided the conditions of Theorem 2.1 are satisfied. The conditions $J \leq K$ and $K \lesssim (n/\log n)^{\delta/(2+\delta)}$ are satisfied by hypothesis.

Corollary 5.1 (under Assumptions 3 and 4 and the fact that $\{(X_i, Y_{2i})\}_{i=1}^n$ are i.i.d. and $J \leq K$) yields

$$\|(\tilde{B}'\tilde{B}/n) - I_K\| = O_p(\sqrt{K(\log K)/n}) \quad (63)$$

$$\|\hat{S} - S\| = O_p(\sqrt{K(\log K)/n}). \quad (64)$$

Therefore, the conditions of Theorem 2.1 are satisfied by these rates and the conditions on J and K in Theorem 3.1.

It remains to control the approximation error $\|P_n h_0 - h_0\|_\infty$. Under Assumptions 1, 3 (with $\Psi_J = \text{BSpl}(J, [0, 1]^d, \gamma)$ or $\text{Wav}(J, [0, 1]^d, \gamma)$) and 5 there exists a $\pi_J h_0 = \tilde{\psi}^{J'} c_J \in \Psi_J$ with $c_J \in \mathbb{R}^J$ such that

$$\|h_0 - \pi_J h_0\|_\infty = O(J^{-p/d}) \quad (65)$$

(see, e.g., Huang (1998)) so it suffices to control $\|P_n h_0 - \pi_J h_0\|_\infty$.

Both $P_n h_0$ and $\pi_J h_0$ lie in Ψ_J , so $\|P_n h_0 - \pi_J h_0\|_\infty$ may be rewritten as

$$\|P_n h_0 - \pi_J h_0\|_\infty = \frac{\|P_n h_0 - \pi_J h_0\|_\infty}{\|\Pi_K T(P_n h_0 - \pi_J h_0)\|_\infty} \times \|\Pi_K T(P_n h_0 - \pi_J h_0)\|_\infty \quad (66)$$

$$\leq \sigma_{\infty, JK}^{-1} \times \|\Pi_K T(P_n h_0 - \pi_J h_0)\|_\infty \quad (67)$$

where

$$\Pi_K T(P_n h_0 - \pi_J h_0)(x) = \tilde{b}^K(x) S' [\hat{S}(\tilde{B}'\tilde{B}/n) - \hat{S}]^{-1} \hat{S}(\tilde{B}'\tilde{B}/n)^{-1} \tilde{B}'(H_0 - \Psi c_J)/n. \quad (68)$$

Define the $K \times K$ matrices

$$\begin{aligned} D &= S'[SS']^{-1}S \\ \hat{D} &= (\tilde{B}'\tilde{B}/n)^{-1} \hat{S}' [\hat{S}(\tilde{B}'\tilde{B}/n) - \hat{S}]^{-1} \hat{S}(\tilde{B}'\tilde{B}/n)^{-1}. \end{aligned} \quad (69)$$

By the triangle inequality,

$$\|\Pi_K T(P_n h_0 - \pi_J h_0)\|_\infty \quad (70)$$

$$\leq \|\tilde{b}^K(x) \hat{D} \tilde{B}'(H_0 - \Psi c_J)/n\|_\infty \quad (71)$$

$$+ \|\tilde{b}^K(x) \{S' - (\tilde{B}'\tilde{B}/n)^{-1} \hat{S}'\} [\hat{S}(\tilde{B}'\tilde{B}/n) - \hat{S}]^{-1} \hat{S}(\tilde{B}'\tilde{B}/n)^{-1} \tilde{B}'(H_0 - \Psi c_J)/n\|_\infty. \quad (72)$$

The arguments below will show that

$$\|\Pi_K T(P_n h_0 - \pi_J h_0)\|_\infty = O_p(\sqrt{K(\log n)/n}) \times \|h_0 - \pi_J h_0\|_\infty + O_p(1) \times \|\Pi_K T(h_0 - \pi_J h_0)\|_\infty. \quad (73)$$

Substituting (73) into (67) and using Assumption 7(ii), the bound $\sigma_{\infty, JK}^{-1} \lesssim \sqrt{J} \times \sigma_{JK}^{-1}$ (under Assumption 3), equation (65), and the condition $p \geq d/2$ in Assumption 5 yields the desired result

$$\|P_n h_0 - \pi_J h_0\|_\infty = O_p(J^{-p/d} + \sigma_{JK}^{-1} \sqrt{K(\log n)/n}). \quad (74)$$

Control of (71): By the triangle and Cauchy-Schwarz inequalities and compatibility of the spectral

norm under multiplication,

$$\begin{aligned}
(71) &\leq \|\tilde{b}^K(x)(\widehat{D} - D)\tilde{B}'(H_0 - \Psi_{c_J})/n\|_\infty \\
&\quad + \|\tilde{b}^K(x)D\{\tilde{B}'(H_0 - \Psi_{c_J})/n - E[\tilde{b}^K(X_i)(h_0(Y_{2i}) - \pi_J h_0(Y_{2i}))]\}\|_\infty \\
&\quad + \|\tilde{b}^K(x)DE[\tilde{b}^K(X_i)\Pi_K T(h_0 - \pi_J h_0)(X_i)]\|_\infty
\end{aligned} \tag{75}$$

$$\begin{aligned}
&\lesssim \sqrt{K}\|\widehat{D} - D\|\{O_p(\sqrt{K/n}) \times \|h_0 - \pi_J h_0\|_\infty + \|\Pi_K T(h_0 - \pi_J h_0)\|_{L^2(X)}\} \\
&\quad + O_p(\sqrt{K(\log n)/n}) \times \|h_0 - \pi_J h_0\|_\infty \\
&\quad + \|\tilde{b}^K(x)DE[\tilde{b}^K(X_i)\Pi_K T(h_0 - \pi_J h_0)(X_i)]\|_\infty
\end{aligned} \tag{76}$$

by Lemma C.3 and properties of the spectral norm. Lemma C.2 (the conditions of Lemma C.2 are satisfied by virtue of (63) and (64) and the condition $\sigma_{JK}^{-1}K\sqrt{(\log n)/n} \lesssim 1$) and the condition $\sigma_{JK}^{-1}K\sqrt{(\log n)/n} \lesssim 1$ yield $\sqrt{K}\|\widehat{D} - D\| = O_p(1)$. Finally, Lemma C.1 (under Assumptions 1 and 4) provides that

$$\|\tilde{b}^K(x)DE[\tilde{b}^K(X_i)\Pi_K T(h_0 - \pi_J h_0)(X_i)]\|_\infty \lesssim \|\Pi_K T(h_0 - \pi_J h_0)\|_\infty \tag{77}$$

and so

$$(71) = O_p(\sqrt{K(\log n)/n}) \times \|h_0 - \pi_J h_0\|_\infty + O_p(1) \times \|\Pi_K T(h_0 - \pi_J h_0)\|_\infty \tag{78}$$

as required.

Control of (72): By the Cauchy-Schwarz inequality, compatibility of the spectral norm under multiplication, and Assumption 4,

$$(72) \lesssim \sqrt{K}\|S' - (\tilde{B}'\tilde{B}/n)^{-\widehat{S}}\| \|\widehat{S}(\tilde{B}'\tilde{B}/n)^{-\widehat{S}} - \widehat{S}(\tilde{B}'\tilde{B}/n)^-\| \|\tilde{B}'(H_0 - \Psi_{c_J})/n\| \tag{79}$$

$$\lesssim \sigma_{JK}^{-1}\sqrt{K}\{\|(\tilde{B}'\tilde{B}/n) - I_K\| + \|\widehat{S} - S\|\} \|\tilde{B}'(H_0 - \Psi_{c_J})/n\| \tag{80}$$

where the second line holds wpa1, by Lemma C.2 (the conditions of Lemma C.2 are satisfied by virtue of (63) and (64) and the condition $\sigma_{JK}^{-1}K\sqrt{(\log n)/n} \lesssim 1$). Applying (63) and (64) and the condition $\sigma_{JK}^{-1}K\sqrt{(\log n)/n} \lesssim 1$ again yields

$$(72) = O_p(1) \times \|\tilde{B}'(H_0 - \Psi_{c_J})/n\| \tag{81}$$

$$= O_p(\sqrt{K/n}) \times \|h_0 - \pi_J h_0\|_\infty + O_p(1) \times \|\Pi_K T(h_0 - \pi_J h_0)\|_\infty \tag{82}$$

by Lemma C.3, equation (65), and the relation between $\|\cdot\|_\infty$ and $\|\cdot\|_{L^2(X)}$. ■

Proof of Lemma 3.1. As already mentioned, Assumption 7(i) implies that the operators T , T^* , TT^* and T^*T are all compact with the singular value system $\{\mu_k; \phi_{1k}, \phi_{0k}\}_{k=1}^\infty$ where $\mu_1 = 1 \geq \mu_2 \geq \mu_3 \geq \dots \searrow 0$. For any $h \in B(p, L) \subset L^2(Y_2)$, $g \in L^2(X)$, we have

$$(Th)(x) = \sum_{k=1}^\infty \mu_k \langle h, \phi_{1k} \rangle_{Y_2} \phi_{0k}(x), \quad (T^*g)(y_2) = \sum_{k=1}^\infty \mu_k \langle g, \phi_{0k} \rangle_X \phi_{1k}(y_2).$$

Let $\mathcal{P}_J = \text{clsp}\{\phi_{0k} : k = 1, \dots, J\}$, and note that \mathcal{P}_J is a closed linear subspace of B_K under the conditions of part (3).

To prove part (3), for any $h \in \Psi_J$ with $J \leq K$, we have

$$\Pi_K Th(\cdot) = \sum_{j=1}^K \langle Th, \tilde{b}_{Kj} \rangle_X \tilde{b}_{Kj}(\cdot) \quad (83)$$

$$= \sum_{j=1}^J \langle Th, \phi_{0j} \rangle_X \phi_{0j}(\cdot) + R(\cdot, h) \quad (84)$$

for some $R(\cdot, h) \in B_K \setminus \mathcal{P}_J$. Therefore

$$\sigma_{JK}^2 = \inf_{h \in \Psi_J: \|h\|_{L^2(\mathcal{Y}_2)}=1} \|\Pi_K Th(X)\|_{L^2(X)}^2 \quad (85)$$

$$= \inf_{h \in \Psi_J: \|h\|_{L^2(\mathcal{Y}_2)}=1} \left\| \sum_{j=1}^J \langle Th, \phi_{0j} \rangle_X \phi_{0j}(\cdot) + R(\cdot, h) \right\|_{L^2(X)}^2 \quad (86)$$

$$= \inf_{h \in \Psi_J: \|h\|_{L^2(\mathcal{Y}_2)}=1} \left(\sum_{j=1}^J \langle Th, \phi_{0j} \rangle_X^2 + \|R(\cdot, h)\|_{L^2(X)}^2 \right) \quad (87)$$

$$\geq \inf_{h \in \Psi_J: \|h\|_{L^2(\mathcal{Y}_2)}=1} \left(\sum_{j=1}^J \langle Th, \phi_{0j} \rangle_X^2 \right) \quad (88)$$

$$= \inf_{h \in \Psi_J: \|h\|_{L^2(\mathcal{Y}_2)}=1} \left(\sum_{j=1}^J \mu_j^2 \langle h, \phi_{1j} \rangle_{\mathcal{Y}_2}^2 \right) \quad (89)$$

$$\geq \mu_J^2 \inf_{h \in \Psi_J: \|h\|_{L^2(\mathcal{Y}_2)}=1} \left(\sum_{j=1}^J \langle h, \phi_{1j} \rangle_{\mathcal{Y}_2}^2 \right) = \mu_J^2. \quad (90)$$

This, together with part (1), gives $1/\mu_J \geq \sigma_{JK}^{-1} \geq \tau_{2,2,J} \geq 1/\mu_J$. ■

Proof of Theorem 3.3. Our proof proceeds by application of Theorem 2.5 of Tsybakov (2009) (page 99).

We first explain the scalar ($d = 1$) case in detail. Let $\{\phi_{jk}, \psi_{jk}\}_{j,k}$ be a wavelet basis for $L^2([0, 1])$ as in the construction of Cohen et al. (1993a,b) with regularity $\gamma > \max\{p, \varsigma\}$ using a pair (ϕ, ψ) for which $\text{support}(\phi) = \text{support}(\psi) = [0, 2N - 1]$. The precise type of wavelet is not important, but we do require that $\|\psi\|_\infty < \infty$. For given j , the wavelet space W_j consists of 2^j functions $\{\psi_{jk}\}_{1 \leq k \leq 2^j}$, such that $\{\psi_{jk}\}_{1 \leq k \leq 2^j - 2N}$ are interior wavelets for which $\psi_{jk}(\cdot) = 2^{j/2} \psi(2^j(\cdot) - k)$. We will choose j deterministically with n , such that

$$2^j \asymp (n/\log n)^{1/(2(p+\varsigma)+1)}. \quad (91)$$

By construction, the support of each interior wavelet is an interval of length $2^{-j}(2N - 1)$. Thus for all j sufficiently large,⁷ we may choose a set $M \subset \{1, \dots, 2^j - 2N\}$ of interior wavelets with $\#(M) \gtrsim 2^j$ such that $\text{support}(\psi_{jm}) \cap \text{support}(\psi_{jm'}) = \emptyset$ for all $m, m' \in M$ with $m \neq m'$. Note also that by construction we have $\#(M) \leq 2^j$ (since there are $2^j - 2N$ interior wavelets).

We begin by defining a family of submodels. Let $h_0 \in B(p, L)$ be such that $\|h_0\|_{B_{\infty, \infty}^p(\mathcal{Y}_2)} \leq L/2$, and for each $m \in M$ let

$$h_m = h_0 + c_0 2^{-j(p+1/2)} \psi_{jm} \quad (92)$$

⁷Hence the lim inf in our statement of the Lemma.

where c_0 is a positive constant to be defined subsequently. Noting that

$$c_0 2^{-j(p+1/2)} \|\psi_{jm}\|_{B_{\infty,\infty}^p} \lesssim c_0 2^{-j(p+1/2)} \|\psi_{jm}\|_{b_{\infty,\infty}^p} \quad (93)$$

$$\leq c_0 \quad (94)$$

it follows by the triangle inequality that $\|h_m\|_{B_{\infty,\infty}^p} \leq L$ uniformly in m for all sufficiently small c_0 . For $m \in \{0\} \cup M$ let P_m be the joint distribution of $\{(X_i, Y_{1i})\}_{i=1}^n$ with $Y_{1i} = Th_m(X_i) + u_i$ for the Gaussian NPIR model (20).

To apply Theorem 2.5 of Tsybakov (2009), first note that for any $m \in M$

$$\|h_0 - h_m\|_{\infty} = c_0 2^{-j(p+1/2)} \|\psi_{jm}\|_{\infty} \quad (95)$$

$$= c_0 2^{-jp} \|\psi\|_{\infty} \quad (96)$$

and for any $m, m' \in M$ with $m \neq m'$

$$\|h_m - h_{m'}\|_{\infty} = c_0 2^{-j(p+1/2)} \|\psi_{jm} - \psi_{jm'}\|_{\infty} \quad (97)$$

$$= 2c_0 2^{-jp} \|\psi\|_{\infty} \quad (98)$$

by virtue of the disjoint support of $\{\psi_{jm}\}_{m \in M}$. Using the KL divergence for the multivariate normal distribution (under the Gaussian NPIR model (20)), Assumption 8 and the equivalence between the Besov function-space and sequence-space norms, the KL distance $K(P_m, P_0)$ is

$$K(P_m, P_0) \leq \frac{1}{2} \sum_{i=1}^n (c_0 2^{-j(p+1/2)})^2 E \left[\frac{(T\psi_{jm}(X_i))^2}{\sigma^2(X_i)} \right] \quad (99)$$

$$\leq \frac{1}{2} \sum_{i=1}^n (c_0 2^{-j(p+1/2)})^2 \frac{E[(T\psi_{jm}(X_i))^2]}{\sigma_0^2} \quad (100)$$

$$= \frac{n}{2\sigma_0^2} (c_0 2^{-j(p+1/2)})^2 \|(T^*T)^{1/2} \psi_{jm}(Y_2)\|_{L^2(Y_2)}^2 \quad (101)$$

$$\lesssim \frac{n}{2\sigma_0^2} (c_0 2^{-j(p+1/2)})^2 (2^{-j\varsigma})^2 \quad (102)$$

$$= \frac{n}{2\sigma_0^2} c_0^2 2^{-j(2(p+\varsigma)+1)} \quad (103)$$

$$\lesssim c_0^2 \log n \quad (104)$$

since $2^{-j} \asymp ((\log n)/n)^{1/(2(p+\varsigma)+1)}$. Moreover, since $\#(M) \asymp 2^j$, we have

$$\log(\#(M)) \lesssim \log n + \log \log n \quad (105)$$

Therefore, we may choose c_0 sufficiently small that $\|h_m\|_{B_{\infty,\infty}^p} \leq L$ and $K(P_m, P_0) \leq \frac{1}{8} \log(\#(M))$ uniformly in m for all n sufficiently large. All conditions of Theorem 2.5 of Tsybakov (2009) are satisfied and hence we obtain the lower bound result.

The multivariate case uses similar arguments for a tensor-product wavelet basis (see Triebel (2006, 2008)). We choose the same j for each univariate spaces such that $2^j \asymp (n/\log n)^{1/(2(p+\varsigma)+d)}$ and so the tensor-product wavelet space has dimension $2^{jd} \asymp (n/\log n)^{d/(2(p+\varsigma)+d)}$. We construct the same family of submodels,

setting $h_m = h_0 + c_0 2^{-j(p+d/2)} \psi_{jm}$ where ψ_{jm} is now the product of d interior univariate wavelets defined previously. Since we take the product of d univariate interior wavelets, we again obtain

$$\|h_m - h_{m'}\|_\infty \gtrsim c_0 2^{-jp} \quad (106)$$

for each $m, m' \in \{0\} \cup M$ with $m \neq m'$, and

$$K(P_m, P_0) \lesssim \frac{n}{2\sigma_0^2} (c_0 2^{-j(p+d/2)})^2 (2^{-j\varsigma})^2 \quad (107)$$

$$= \frac{n}{2\sigma_0^2} c_0^2 2^{-j(2(p+\varsigma)+d)} \quad (108)$$

$$\lesssim c_0^2 \log n. \quad (109)$$

The result follows as in the univariate case. ■

B.3 Proofs for Section 4

Proof of Theorem 4.1. The variance term is immediate from Theorem 2.1 with $\sigma_{JK} = 1$. The bias calculation follows from Huang (2003) under Assumptions 1(ii), 4 (with $B_K = \text{BSpl}(K, [0, 1]^d, \gamma)$ or $\text{Wav}(K, [0, 1]^d, \gamma)$), and 5 and the fact that the empirical and true $L^2(X)$ norms are equivalent over B_K wpa1 by virtue of the condition $\|\tilde{B}'\tilde{B}/n - I_K\| = o_p(1)$, which is implied by Condition (ii). ■

Proof of Corollary 4.1. By Theorem 4.1, the optimal sup-norm convergence rate $(n/\log n)^{-p/(2p+d)}$ is achieved by setting $K \asymp (n/\log n)^{d/(2p+d)}$, with $\delta \geq d/p$ for condition (i) to hold. (1) When the regressors are i.i.d., by Lemma 5.2, condition (ii) is satisfied provided that $d \leq 2p$ (which is assumed in Assumption 5). (2) When the regressors are exponentially β -mixing, by Lemma 5.3, condition (ii) is satisfied provided that $d < 2p$. (3) When the regressors are algebraically β -mixing at rate γ , by Lemma 5.3, condition (ii) is satisfied provided that $(2 + \gamma)d < 2\gamma p$. ■

B.4 Proofs for Section 5

Proof of Corollary 5.1. Follows by application of Theorem 5.1 with $t = C\sigma_n \sqrt{\log(d_1 + d_2)}$ for sufficiently large C , and applying the condition $R_n \sqrt{\log(d_1 + d_2)} = o(\sigma_n)$. ■

Proof of Theorem 5.2. By Berbee's lemma (enlarging the probability space as necessary) the process $\{X_i\}$ can be coupled with a process X_i^* such that $Y_k := \{X_{(k-1)q+1}, \dots, X_{kq}\}$ and $Y_k^* := \{X_{(k-1)q+1}^*, \dots, X_{kq}^*\}$ are identically distributed for each $k \geq 1$, $\mathbb{P}(Y_k \neq Y_k^*) \leq \beta(q)$ for each $k \geq 1$ and $\{Y_1^*, Y_3^*, \dots\}$ are independent and $\{Y_2^*, Y_4^*, \dots\}$ are independent (see, e.g., Doukhan et al. (1995)). Let I_e and I_o denote the indices of $\{1, \dots, n\}$ corresponding to the odd- and even-numbered blocks, and I_r the indices in the remainder, so $I_r = q[n/q] + 1, \dots, n$ when $q[n/q] < n$ and $I_r = \emptyset$ when $q[n/q] = n$.

Let $\Xi_{i,n}^* = \Xi(X_{i,n}^*)$. By the triangle inequality,

$$\begin{aligned} & \mathbb{P}(\|\sum_{i=1}^n \Xi_{i,n}\| \geq 6t) \\ & \leq \mathbb{P}(\|\sum_{i=1}^{[n/q]q} \Xi_{i,n}^*\| + \|\sum_{i \in I_r} \Xi_{i,n}\| + \|\sum_{i=1}^{[n/q]q} (\Xi_{i,n}^* - \Xi_{i,n})\| \geq 6t) \\ & \leq \frac{n}{q} \beta(q) + \mathbb{P}(\|\sum_{i \in I_r} \Xi_{i,n}\| \geq t) + \mathbb{P}(\|\sum_{i \in I_e} \Xi_{i,n}^*\| \geq t) + \mathbb{P}(\|\sum_{i \in I_o} \Xi_{i,n}^*\| \geq t) \end{aligned} \quad (110)$$

To control the last two terms we apply Theorem 5.1, recognizing that $\sum_{i \in I_e} \Xi_{i,n}^*$ and $\sum_{i \in I_o} \Xi_{i,n}^*$ are each the sum of fewer than $[n/q]$ independent $d_1 \times d_2$ matrices, namely $W_k^* = \sum_{i=(k-1)q+1}^{kq} \Xi_{i,n}^*$. Moreover each W_k^* satisfies $\|W_k^*\| \leq qR_n$ and $\max\{\|E[W_k^* W_k^{*'}]\|, \|E[W_k^{*'} W_k^*]\|\} \leq q^2 s_n$. Theorem 5.1 then yields

$$\mathbb{P}\left(\left\|\sum_{i \in I_e} \Xi_{i,n}^*\right\| \geq t\right) \leq (d_1 + d_2) \exp\left(\frac{-t^2/2}{nqs_n^2 + qR_n t/3}\right) \quad (111)$$

and similarly for I_o . ■

Proof of Corollary 5.2. Follows by application of Theorem 5.2 with $t = Cs_n \sqrt{nq \log(d_1 + d_2)}$ for sufficiently large C , and applying the conditions $\frac{n}{q} \beta(q) = o(1)$ and $R_n \sqrt{q \log(d_1 + d_2)} = o(s_n \sqrt{n})$. ■

Proof of Lemma 5.1. Let $G_K = E[b^K(X)b^K(X)']$. Since $B_K = \text{clsp}\{b_1, \dots, b_K\}$, we have:

$$\begin{aligned} & \sup\{|\frac{1}{n} \sum_{i=1}^n b(X_i)^2 - 1| : b \in B_K, E[b(X)^2] = 1\} \\ &= \sup\{|c'(B'B/n - G_K)c| : c \in \mathbb{R}^K, \|G_K^{1/2}c\| = 1\} \end{aligned} \quad (112)$$

$$= \sup\{|c'G_K^{1/2}(G_K^{-1/2}(B'B/n)G_K^{-1/2} - I_K)G_K^{1/2}c| : c \in \mathbb{R}^K, \|G_K^{1/2}c\| = 1\} \quad (113)$$

$$= \sup\{|c'(\tilde{B}'\tilde{B}/n - I_K)c| : c \in \mathbb{R}^K, \|c\| = 1\} \quad (114)$$

$$= \|\tilde{B}'\tilde{B}/n - I_K\|_2^2 \quad (115)$$

as required. ■

Proof of Lemma 5.2. Follows by application of Corollary 5.1 with $\Xi_{i,n} = n^{-1}(\tilde{b}^K(X_i)\tilde{b}^K(X_i)' - I_K)$, $R_n \lesssim n^{-1}(\zeta_0(K)^2 + 1)$, and $\sigma_n^2 \lesssim n^{-1}(\zeta_0(K)^2 + 1)$. ■

Proof of Lemma 5.3. Follows by application of Corollary 5.2 with $\Xi_{i,n} = n^{-1}(\tilde{b}^K(X_i)\tilde{b}^K(X_i)' - I_K)$, $R_n \lesssim n^{-1}(\zeta_0(K)^2 + 1)$, and $s_n^2 \lesssim n^{-2}(\zeta_0(K)^2 + 1)$. ■

C Supplementary lemmas and their proofs

Huang (2003) provides conditions under which the operator norm of orthogonal projections onto sieve spaces are *stable* in sup norm as the dimension of the sieve space increases. The following Lemma shows the same is true for the operator $Q_K : L^\infty(X) \rightarrow L^\infty(X)$ given by

$$Q_K u(x) = \tilde{b}^K(x) D E[\tilde{b}^K(X)u(X)] \quad (116)$$

where $D = S'[SS']^{-1}S$, i.e.

$$\limsup_{K \rightarrow \infty} \sup_{u \in L^\infty(X)} \frac{\|Q_K u\|_\infty}{\|u\|_\infty} \leq C \quad (117)$$

for some finite positive constant C . The proof follows by simple modification of the arguments in Theorem A.1 in Huang (2003) (see also Corollary A.1 of Huang (2003)).

Lemma C.1 Q_K is stable in sup norm under Assumption 1 and 4.

Proof of Lemma C.1. The assumptions of Theorem A.1 of Huang (2003) with ν and ν_n taken to be the distribution of X are satisfied under Assumption 1. Let P_K denote the orthogonal projection onto the sieve space, i.e.

$$P_K u(x) = b^K(x)' E[b^K(X)b^K(X)']^{-1} E[b^K(X)u(X)] \quad (118)$$

for any $u \in L^\infty(X)$. Let $\langle \cdot, \cdot \rangle$ denote the $L^2(X)$ inner product. Since D is an orthogonal projection matrix and P_K is a $L^2(X)$ orthogonal projection onto B_K , for any $u \in L^\infty(X)$

$$\begin{aligned} \|Q_K u\|_{L^2(X)}^2 &= E[u(X)\tilde{b}^K(X)']D^2E[\tilde{b}^K(X)u(X)] \\ &\leq E[u(X)\tilde{b}^K(X)']E[\tilde{b}^K(X)u(X)] \\ &= \|P_K u\|_{L^2(X)}^2 \\ &\leq \|u\|_{L^2(X)}^2. \end{aligned} \tag{119}$$

As in Huang (2003), let Δ index a partition of \mathcal{X} into finitely many polyhedra. Let $v \in L^\infty(X)$ be supported on δ_0 for some $\delta_0 \in \Delta$ (i.e. $v(x) = 0$ if $x \notin \delta_0$). For some coefficients $\alpha_1, \dots, \alpha_K$,

$$Q_K v(x) = \sum_{i=1}^K \alpha_i b_{Ki}(x). \tag{120}$$

Let $d(\cdot, \cdot)$ be the distance measure between elements of Δ defined in the Appendix of Huang (2003). Let l be a nonnegative integer and let $I_l \subset \{1, \dots, K\}$ be the set of indices such that for any $i \in I_l$ the basis function b_{Ki} is active on a $\delta \in \Delta$ with $d(\delta, \delta_0) \leq l$. Finally, let

$$v_l(x) = \sum_{i \in I_l} \alpha_i b_{Ki}(x). \tag{121}$$

For any $v \in L^\infty(X)$,

$$\begin{aligned} &\|Q_K v\|_{L^2(X)}^2 \\ &= \|Q_K v - v\|_{L^2(X)}^2 + \|v\|_{L^2(X)}^2 + 2\langle Q_K v - v, v \rangle \end{aligned} \tag{122}$$

$$= \|P_K v - v\|_{L^2(X)}^2 + \|Q_K v - P_K v\|_{L^2(X)}^2 + 2\langle Q_K v - P_K v, P_K v - v \rangle + \|v\|_{L^2(X)}^2 + 2\langle Q_K v - v, v \rangle \tag{123}$$

$$\leq \|v_l - v\|_{L^2(X)}^2 + \|Q_K v - P_K v\|_{L^2(X)}^2 + 2\langle Q_K v - P_K v, P_K v - v \rangle + \|v\|_{L^2(X)}^2 + 2\langle Q_K v - v, v \rangle \tag{124}$$

$$= \|v_l - v\|_{L^2(X)}^2 + \|v\|_{L^2(X)}^2 + \langle Q_K v - P_K v, Q_K v + P_K v - 2v \rangle + 2\langle Q_K v - v, v \rangle \tag{125}$$

$$= \|v_l - v\|_{L^2(X)}^2 + \|v\|_{L^2(X)}^2 + \|Q_K v\|_{L^2(X)}^2 - \|P_K v\|_{L^2(X)}^2 - 2\langle Q_K v - P_K v, v \rangle + 2\langle Q_K v - v, v \rangle \tag{126}$$

$$= \|v_l - v\|_{L^2(X)}^2 + \|v\|_{L^2(X)}^2 + \|Q_K v\|_{L^2(X)}^2 - \|P_K v\|_{L^2(X)}^2 + 2\langle P_K v - v, v \rangle \tag{127}$$

$$= \|v_l - v\|_{L^2(X)}^2 + \|v\|_{L^2(X)}^2 + \|Q_K v\|_{L^2(X)}^2 + \|P_K v\|_{L^2(X)}^2 - 2\|v\|_{L^2(X)}^2 \tag{128}$$

$$\leq \|v_l - v\|_{L^2(X)}^2 + \|v\|_{L^2(X)}^2 \tag{129}$$

where (124) uses the fact that P_K is an orthogonal projection, and (129) follows from (119). The remainder of the proof of Theorem A.1 of Huang (2003) goes through under these modifications. ■

The next lemma provides useful bounds on the estimated matrices encountered in the body of the paper. Recall the definitions of \hat{D} and D in expression (69).

Lemma C.2 *Under Assumption 3(ii) and 4(ii), if $J \leq K$, $\|(\tilde{B}'\tilde{B}/n) - I_K\| = o_p(1)$, then wpa1*

(i) $(\tilde{B}'\tilde{B}/n)$ is invertible and $\|(\tilde{B}'\tilde{B}/n)^{-1}\| \leq 2$

(ii) $\|(\tilde{B}'\tilde{B}/n)^{-1}\hat{S}' - S'\| \lesssim \|(\tilde{B}'\tilde{B}/n) - I_K\| + \|\hat{S} - S\|$

$$(iii) \quad \|(\tilde{B}'\tilde{B}/n)^{-1/2}\hat{S}' - S'\| \lesssim \|(\tilde{B}'\tilde{B}/n) - I_K\| + \|\hat{S} - S\|.$$

If, in addition, $\sigma_{JK}^{-1}(\|(\tilde{B}'\tilde{B}/n) - I_K\| + \|\hat{S} - S\|) = o_p(1)$, then wpa1

$$(iv) \quad (\tilde{B}'\tilde{B}/n)^{-1/2}\hat{S}' \text{ has full column rank and } \hat{S}(\tilde{B}'\tilde{B}/n)^{-1}\hat{S} \text{ is invertible}$$

$$(v) \quad \|\hat{D} - D\| \lesssim \|(\tilde{B}'\tilde{B}/n) - I_K\| + \sigma_{JK}^{-1}(\|\hat{S} - S\| + \|(\tilde{B}'\tilde{B}/n) - I_K\|)$$

$$(vi) \quad \|[\hat{S}(\tilde{B}'\tilde{B}/n)^{-1}\hat{S}' - \hat{S}(\tilde{B}'\tilde{B}/n)^{-1} - [SS']^{-1}S]\| \lesssim \sigma_{JK}^{-1}\|(\tilde{B}'\tilde{B}/n) - I_K\| + \sigma_{JK}^{-2}(\|\hat{S} - S\| + \|(\tilde{B}'\tilde{B}/n) - I_K\|).$$

Proof of Lemma C.2. We prove Lemma C.2 by part. Note that under Assumption 3(ii) and 4(ii), $\|S\| \leq 1$ since S is isomorphic to the $L^2(X)$ orthogonal projection of T onto the space B_K , restricted to Ψ_J .

(i) Let \mathcal{A}_n denote the event $\|(\tilde{B}'\tilde{B}/n) - I_K\| \leq \frac{1}{2}$. The condition $\|(\tilde{B}'\tilde{B}/n) - I_K\| = o_p(1)$ implies that $\mathbb{P}(\mathcal{A}_n^c) = o(1)$. Clearly $\|(\tilde{B}'\tilde{B}/n)^{-1}\| \leq 2$ on \mathcal{A}_n .

(ii) Working on \mathcal{A}_n (so we replace the generalized inverse with an inverse), Assumption 4(ii), the triangle inequality, and compatibility of the spectral norm under multiplication yields

$$\|(\tilde{B}'\tilde{B}/n)^{-1}\hat{S}' - S'\| \leq \|(\tilde{B}'\tilde{B}/n)^{-1}\hat{S}' - (\tilde{B}'\tilde{B}/n)^{-1}S'\| + \|(\tilde{B}'\tilde{B}/n)^{-1}S' - S'\| \quad (130)$$

$$\leq \|(\tilde{B}'\tilde{B}/n)^{-1}\|\|\hat{S} - S\| + \|(\tilde{B}'\tilde{B}/n)^{-1} - I_K\|\|S'\| \quad (131)$$

$$\leq 2\|\hat{S} - S\| + \|(\tilde{B}'\tilde{B}/n)^{-1} - I_K\| \quad (132)$$

$$= 2\|\hat{S} - S\| + \|(\tilde{B}'\tilde{B}/n)^{-1}[(\tilde{B}'\tilde{B}/n) - I_K]\| \quad (133)$$

$$\leq 2\|\hat{S} - S\| + 2\|(\tilde{B}'\tilde{B}/n) - I_K\| \quad (134)$$

(iii) Follows the same arguments as (ii), noting additionally that $\lambda_{\min}((\tilde{B}'\tilde{B}/n)^{-1}) \leq \frac{1}{2}$ on \mathcal{A}_n , in which case

$$\|(\tilde{B}'\tilde{B}/n)^{-1/2} - I_K\| \leq (1 + 2^{-1/2})^{-1}\|(\tilde{B}'\tilde{B}/n) - I_K\| \quad (135)$$

by Lemma 2.2 of Schmitt (1992).

(iv) Let $s_J(A)$ denote the J th largest singular value of a $J \times K$ matrix A . Weyl's inequality yields

$$|s_J(\hat{S}(\tilde{B}'\tilde{B}/n)^{-1/2}) - \sigma_{JK}| \leq \|(\tilde{B}'\tilde{B}/n)^{-1/2}\hat{S}' - S'\|. \quad (136)$$

This and the condition $\sigma_{JK}^{-1}(\|(\tilde{B}'\tilde{B}/n) - I_K\| + \|\hat{S} - S\|) = o_p(1)$ together imply that

$$|s_J(\hat{S}(\tilde{B}'\tilde{B}/n)^{-1/2}) - \sigma_{JK}| \leq \frac{1}{2}\sigma_{JK} \quad (137)$$

wpa1. Let \mathcal{C}_n be the intersection of \mathcal{A}_n with the set on which this bound obtains. Then $\mathbb{P}(\mathcal{C}_n^c) = o(1)$. Clearly $(\tilde{B}'\tilde{B}/n)^{-1/2}\hat{S}'$ has full column rank J and $\hat{S}(\tilde{B}'\tilde{B}/n)^{-1}\hat{S}$ is invertible on \mathcal{C}_n .

(v) On $\mathcal{C}_n \subseteq \mathcal{A}_n$ we have $\|(\tilde{B}'\tilde{B}/n)^{-1/2}\| \leq \sqrt{2}$. Working on \mathcal{C}_n , similar arguments to those used to prove parts (ii) and (iii) yield

$$\|\hat{D} - D\| \leq \|(\tilde{B}'\tilde{B}/n)^{-1/2} - I_K\|(\|(\tilde{B}'\tilde{B}/n)^{-1/2}\| + 1) + \|\hat{Q} - Q\|\|(\tilde{B}'\tilde{B}/n)^{-1/2}\| \quad (138)$$

$$\leq (1 + \sqrt{2})\|(\tilde{B}'\tilde{B}/n)^{-1/2} - I_K\| + \sqrt{2}\|\hat{Q} - Q\|. \quad (139)$$

Since \widehat{Q} and Q are orthogonal projection matrices, part (1.5) of Theorem 1.1 of Li et al. (2013) implies

$$\|\widehat{Q} - Q\| \leq \sigma_{JK}^{-1} \|(\widetilde{B}'\widetilde{B}/n)^{-1/2}\widehat{S}' - S'\| \quad (140)$$

on \mathcal{C}_n . Part (v) is then proved by substituting (140) and (135) into (139).

(vi) Working on \mathcal{C}_n (so we replace the generalized inverses with inverses), similar arguments used to prove part (v) yield

$$\begin{aligned} & \|[\widehat{S}(\widetilde{B}'\widetilde{B}/n)^{-1}\widehat{S}']^{-1}\widehat{S}(\widetilde{B}'\widetilde{B}/n)^{-1} - [SS']^{-1}S\| \\ \leq & \|[\widehat{S}(\widetilde{B}'\widetilde{B}/n)^{-1}\widehat{S}']^{-1}\widehat{S}(\widetilde{B}'\widetilde{B}/n)^{-1/2}\| \|(\widetilde{B}'\widetilde{B}/n)^{-1/2} - I_K\| \end{aligned} \quad (141)$$

$$\begin{aligned} & + \|[\widehat{S}(\widetilde{B}'\widetilde{B}/n)^{-1}\widehat{S}']^{-1}\widehat{S}(\widetilde{B}'\widetilde{B}/n)^{-1/2} - [SS']^{-1}S\| \\ \leq & 2\sigma_{JK}^{-1} \|(\widetilde{B}'\widetilde{B}/n)^{-1/2} - I_K\| + \|[\widehat{S}(\widetilde{B}'\widetilde{B}/n)^{-1}\widehat{S}']^{-1}\widehat{S}(\widetilde{B}'\widetilde{B}/n)^{-1/2} - [SS']^{-1}S\|. \end{aligned} \quad (142)$$

Theorem 3.1 of Ding and Huang (1997) yields

$$\|[\widehat{S}(\widetilde{B}'\widetilde{B}/n)^{-1}\widehat{S}']^{-1}\widehat{S}(\widetilde{B}'\widetilde{B}/n)^{-1/2} - [SS']^{-1}S\| \lesssim \sigma_{JK}^{-2} \|(\widetilde{B}'\widetilde{B}/n)^{-1/2}\widehat{S}' - S'\| \quad (143)$$

wpa1 by virtue of part (iii) and the condition $\sigma_{JK}^{-1}(\|(\widetilde{B}'\widetilde{B}/n) - I_K\| + \|\widehat{S} - S\|) = o_p(1)$. Substituting (143) and (135) into (142) establishes (vi).

This completes the proof. ■

Lemma C.3 *Under Assumption 4, if $\{X_i, Y_{2i}\}_{i=1}^n$ are i.i.d. then*

$$(i) \quad \|\widetilde{B}'(H_0 - \Psi_{c_J})/n\| \leq O_p(\sqrt{K/n}) \times \|h_0 - \pi_J h_0\|_\infty + \|\Pi_K T(h_0 - \pi_J h_0)\|_{L^2(X)}$$

$$(ii) \quad \|\widetilde{b}^K(x) D\{\widetilde{B}'(H_0 - \Psi_{c_J})/n - E[\widetilde{b}^K(X_i)(h_0(Y_{2i}) - \pi_J h_0(Y_{2i}))]\}\|_\infty = O_p(\sqrt{K(\log n)/n}) \times \|h_0 - \pi_J h_0\|_\infty.$$

Proof of Lemma C.3. We prove Lemma C.3 by part.

(i) First write

$$\begin{aligned} \|\widetilde{B}'(H_0 - \Psi_{c_J})/n\| & \leq \|\widetilde{B}'(H_0 - \Psi_{c_J})/n - E[\widetilde{b}^K(X_i)(h_0(Y_{2i}) - \pi_J h_0(Y_{2i}))]\| \\ & \quad + \|E[\widetilde{b}^K(X_i)(h_0(Y_{2i}) - \pi_J h_0(Y_{2i}))]\| \end{aligned} \quad (144)$$

and note that

$$\|E[\widetilde{b}^K(X_i)(h_0(Y_{2i}) - \pi_J h_0(Y_{2i}))]\|^2 = \|\Pi_K T(h_0 - \pi_J h_0)\|_{L^2(X)}^2. \quad (145)$$

Finally,

$$\|\widetilde{B}'(H_0 - \Psi_{c_J})/n - E[\widetilde{b}^K(X_i)(h_0(Y_{2i}) - \pi_J h_0(Y_{2i}))]\| = O_p(\sqrt{K/n}) \times \|h_0 - \pi_J h_0\|_\infty. \quad (146)$$

by Markov's inequality under Assumption 4 and the fact that $\{X_i, Y_{2i}\}_{i=1}^n$ are i.i.d.

(ii) An argument similar to the proof of Theorem 2.1 converts the problem of controlling the supremum that of controlling the maximum evaluated at finitely many points, where the collection of points has cardinality increasing polynomially in n . Let \mathcal{S}'_n be the set of points. Also define

$$\Delta_{i,J,K} = \tilde{b}^K(X_i)(h_0(Y_{2i}) - \pi_J h_0(Y_{2i})) - E[\tilde{b}^K(X_i)(h_0(Y_{2i}) - \pi_J h_0(Y_{2i}))] \quad (147)$$

Then it suffices to show that sufficiently large C may be chosen that

$$(\#\mathcal{S}'_n) \max_{x_n \in \mathcal{S}'_n} \mathbb{P} \left(\left| \sum_{i=1}^n n^{-1} \tilde{b}^K(x_n) D\Delta_{i,J,K} \right| > C \|h_0 - \pi_J h_0\|_\infty \sqrt{K(\log n)/n} \right) = o(1). \quad (148)$$

The summands in (148) have mean zero (by the law of iterated expectations). Under Assumption 4 the summands in (148) are bounded uniformly for $x_n \in \mathcal{S}'_n$ by

$$|n^{-1} \tilde{b}^K(x_n) D\Delta_{i,J,K}| \lesssim \frac{K}{n} \|h_0 - \pi_J h_0\|_\infty \quad (149)$$

and have variance bounded uniformly for $x_n \in \mathcal{S}'_n$ by

$$E[(n^{-1} \tilde{b}^K(x_n) D\Delta_{i,J,K})^2] \leq \|h_0 - \pi_J h_0\|_\infty^2 \times n^{-2} E[\tilde{b}^K(x_n)' D\tilde{b}^K(X_i) \tilde{b}^K(X_i)' D\tilde{b}^K(x_n)] \quad (150)$$

$$\lesssim \|h_0 - \pi_J h_0\|_\infty^2 \times \frac{K}{n^2}. \quad (151)$$

The result follows for large enough C by Bernstein's inequality for i.i.d. sequences using the bounds (149) and (151).

This completes the proof. ■

References

- AI, C. AND X. CHEN (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843.
- ANDREWS, D. W. K. (2011): "Examples of L2-Complete and Boundedly-Complete Distributions," *Cowles Foundation Discussion Paper No. 1801*.
- BELLONI, A., V. CHERNOZHUKOV, AND K. KATO (2012): "On the Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," Preprint, arXiv:1212.0442v1 [stat.ME].
- BLUNDELL, R., X. CHEN, AND D. KRISTENSEN (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75, 1613–1669.
- BLUNDELL, R. AND J. L. POWELL (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics*, ed. by M. Dewatripont, L. P. Hansen, and S. J. Turnovsky, Cambridge University Press, Cambridge.

- CARROLL, R. J. AND P. HALL (1988): “Optimal Rates of Convergence for Deconvolving a Density,” *Journal of the American Statistical Association*, 83, 1184–1186.
- CATTANEO, M. D. AND M. H. FARRELL (2013): “Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators,” *Journal of Econometrics*, 174, 127–143.
- CAVALIER, L. (2008): “Nonparametric Statistical Inverse Problems,” *Inverse Problems*, 24, 034004.
- CAVALIER, L., G. K. GOLUBEV, D. PICARD, AND A. B. TSYBAKOV (2002): “Oracle Inequalities for Inverse Problems,” *The Annals of Statistics*, 30, 843–874.
- CAVALIER, L. AND N. W. HENGARTNER (2005): “Adaptive Estimation for Inverse Problems with Noisy Operators,” *Inverse Problems*, 21, 1345–1361.
- CHEN, X., V. CHERNOZHUKOV, S. LEE, AND W. K. NEWEY (2013): “Local Identification of Nonparametric and Semiparametric Models,” *Econometrica*, *forthcoming*.
- CHEN, X. AND S. C. LUDVIGSON (2009): “Land of Addicts? An Empirical Investigation of Habit-Based Asset Pricing Models,” *Journal of Applied Econometrics*, 24, 1057–1093.
- CHEN, X. AND D. POUZO (2009): “Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals,” *Journal of Econometrics*, 152, 46–60.
- (2012): “Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals,” *Econometrica*, 80, 277–321.
- CHEN, X. AND M. REISS (2011): “On Rate Optimality for Ill-Posed Inverse Problems in Econometrics,” *Econometric Theory*, 27, 497–521.
- COHEN, A., I. DAUBECHIES, B. JAWERTH, AND P. VIAL (1993a): “Multiresolution Analysis, Wavelets and Fast Algorithms on an Interval,” *Comptes Rendus de l’Académie des Sciences Paris. Série 1, Mathématique*, 316, 417–421.
- COHEN, A., I. DAUBECHIES, AND P. VIAL (1993b): “Wavelets on the Interval and Fast Wavelet Transforms,” *Applied and Computational Harmonic Analysis*, 1, 54–81.
- COHEN, A., M. HOFFMANN, AND M. REISS (2004): “Adaptive Wavelet Galerkin Methods for Linear Inverse Problems,” *SIAM Journal on Numerical Analysis*, 42, 1479–1501.
- DAROLLES, S., Y. FAN, J.-P. FLORENS, AND E. RENAULT (2011): “Nonparametric Instrumental Regression,” *Econometrica*, 79, 1541–1565.
- DE BOOR, C. (2001): *A Practical Guide to Splines*, Springer-Verlag, New York.
- DE JONG, R. M. (2002): “A Note on “Convergence Rates and Asymptotic Normality for Series Estimators”: Uniform Convergence Rates,” *Journal of Econometrics*, 111, 1–9.

- D'HAULTFOEUILLE, X. (2011): "On the Completeness Condition in Nonparametric instrumental regression," *Econometric Theory*, 27, 460–471.
- DING, J. AND L. HUANG (1997): "On the Continuity of Generalized Inverses of Linear Operators in Hilbert Spaces," *Linear Algebra and its Applications*, 262, 229–242.
- DOUKHAN, P., P. MASSART, AND E. RIO (1995): "Invariance Principles for Absolutely Regular Empirical Processes," *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 31, 393–427.
- EFROMOVICH, S. AND V. KOLTCHINSKII (2001): "On Inverse Problems with Unknown Operators," *IEEE Transactions on Information Theory*, 47, 2876–2894.
- FAN, J. (1991): "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *The Annals of Statistics*, 19, 1257–1272.
- 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, 61–75.
- HALL, P. AND J. L. HOROWITZ (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *The Annals of Statistics*, 33, 2904–2929.
- HALL, P. AND A. MEISTER (2007): "A Ridge-Parameter Approach to Deconvolution," *The Annals of Statistics*, 35, 1535–1558.
- HANSEN, B. (2008): "Uniform convergence rates for kernel estimation with dependent data," *Econometric Theory*, 24, 726–748.
- HOFFMANN, M. AND M. REISS (2008): "Nonlinear Estimation for Linear Inverse Problems with Error in the Operator," *The Annals of Statistics*, 36, 310–336.
- HOROWITZ, J. L. (2011): "Applied Nonparametric Instrumental Variables Estimation," *Econometrica*, 79, 347–394.
- HUANG, J. Z. (1998): "Projection Estimation in Multiple Regression with Application to Functional ANOVA Models," *The Annals of Statistics*, 26, 242–272.
- (2003): "Local Asymptotics for Polynomial Spline Regression," *The Annals of Statistics*, 31, 1600–1635.
- JOHNSTONE, I. M. (2013): "Gaussian Estimation: Sequence and Wavelet Models," Manuscript.
- KRESS, R. (1999): *Linear Integral Equations*, Springer-Verlag, Berlin.
- LEE, J. AND P. ROBINSON (2013): "Series Estimation Under Cross-sectional Dependence," Preprint, London School of Economics.

- LI, B., W. LI, AND L. CUI (2013): “New Bounds for Perturbation of the Orthogonal Projection,” *Calcolo*, 50, 69–78.
- LIAO, Y. AND W. JIANG (2011): “Posterior Consistency of Nonparametric Conditional Moment Restricted Models,” *The Annals of Statistics*, 39, 3003–3031.
- LOUBES, J.-M. AND C. MARTEAU (2012): “Adaptive Estimation for an Inverse Regression model with Unknown Operator,” *Statistics & Risk Modeling*, 29, 215–242.
- LOUNICI, K. AND R. NICKL (2011): “Global Uniform Risk Bounds for Wavelet Deconvolution Estimators,” *The Annals of Statistics*, 39, 201–231.
- MASRY, E. (1996): “Multivariate local polynomial regression for time series: uniform strong consistency and rates,” *Journal of Time Series Analysis*, 17, 571–599.
- NEWBY, W. K. (1997): “Convergence Rates and Asymptotic Normality for Series Estimators,” *Journal of Econometrics*, 79, 147–168.
- NEWBY, W. K. AND J. L. POWELL (2003): “Instrumental Variable Estimation of Nonparametric Models,” *Econometrica*, 71, 1565–1578.
- SCHMITT, B. A. (1992): “Perturbation Bounds for Matrix Square Roots and Pythagorean Sums,” *Linear Algebra and its Applications*, 174, 215–227.
- SCHUMACKER, L. L. (2007): *Spline Functions: Basic Theory*, Cambridge University Press, Cambridge.
- SONG, K. (2008): “Uniform Convergence of Series Estimators over Function Spaces,” *Econometric Theory*, 24, 1463–1499.
- STONE, C. J. (1982): “Optimal Global Rates of Convergence for Nonparametric Regression,” *The Annals of Statistics*, 10, 1040–1053.
- TRIEBEL, H. (2006): *Theory of Function Spaces III*, Birkhäuser, Basel.
- (2008): *Function Spaces and Wavelets on Domains*, European Mathematical Society, Zürich.
- TROPP, J. A. (2012): “User-Friendly Tail Bounds for Sums of Random Matrices,” *Foundations of Computational Mathematics*, 12, 389–434.
- TSYBAKOV, A. B. (2009): *Introduction to Nonparametric Estimation*, Springer, New York.
- VAN DE GEER, S. (1995): “Exponential Inequalities for Martingales, with Application to Maximum Likelihood Estimation for Counting Processes,” *The Annals of Statistics*, 23, 1779–1801.
- ZHANG, C.-H. (1990): “Fourier methods for estimating mixing densities and distributions,” *The Annals of Statistics*, 18, 806–831.