Adaptive Estimation and Uniform Confidence Bands for Nonparametric Structural Functions and Elasticities*

Xiaohong Chen† Timothy Christensen‡ Sid Kankanala§

First version: July 24, 2021; Revised version: November 22, 2023

Abstract

We introduce two data-driven procedures for optimal estimation and inference in nonparametric models using instrumental variables. The first is a data-driven choice of sieve dimension for a popular class of sieve two-stage least squares estimators. When implemented with this choice, estimators of both the structural function h_0 and its derivatives (such as elasticities) converge at the fastest possible (i.e., minimax) rates in sup-norm. The second is for constructing uniform confidence bands (UCBs) for h_0 and its derivatives. Our UCBs guarantee coverage over a generic class of data-generating processes and contract at the minimax rate, possibly up to a logarithmic factor. As such, our UCBs are asymptotically more efficient than UCBs based on the usual approach of undersmoothing. As an application, we estimate the elasticity of the intensive margin of firm exports in a monopolistic competition model of international trade. Simulations illustrate the good performance of our procedures in empirically calibrated designs. Our results provide evidence against common parameterizations of the distribution of unobserved firm heterogeneity.

Keywords: Honest and adaptive uniform confidence bands, minimax rate-adaptive estimation, nonparametric instrumental variables, nonparametric estimation of elasticities, international trade.

---

*Authors are in alphabetical order. We are grateful to Francesca Molinari, two anonymous referees, Richard Nickl and Yixiao Sun for helpful suggestions, and to Rodrigo Adao, Costas Arkolakis and Sharat Ganapati for sharing their data. We thank participants of numerous workshops and the 2022 IAAE, 2022 AMES China, KEA2022, 2022 Toulouse Conference on Estimation and Inference in Econometric Models, and 2022 CIREQ Montreal Econometrics Conferences for comments. The data-driven choice of sieve dimension in this paper is based on and supersedes Section 3 of the preprint `arXiv:1508.03365v1` (Chen and Christensen, 2015a). The research is partially supported by the Cowles Foundation Research Funds (Chen) and the National Science Foundation under Grant No. SES-1919034 (Christensen).

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

‡Department of Economics, University College London. `t.christensen@ucl.ac.uk`

§Department of Economics, Yale University. `sid.kankanala@yale.edu`

1 Introduction

With easier access to large data sets, there is increasing interest in estimating flexible, nonparametric structural functions and their derivatives, such as elasticities or other marginal effects. In many applications, the structural function h_0 is identified by a conditional moment restriction

$$\mathbb{E}[Y - h_0(X)|W] = 0 \text{ (almost surely)}, \tag{1}$$

where Y (a scalar) and/or some elements of X (a vector) are endogenous, W is a vector of instrumental variables, and the conditional distribution of (X, Y) given W is otherwise unspecified. Examples include consumer demand (Blundell, Chen, and Kristensen, 2007; Blundell, Horowitz, and Parey, 2017), demand for differentiated products (Berry and Haile, 2014; Compiani, 2022), and international trade (Adao, Costinot, and Donaldson, 2017; Adao, Arkolakis, and Ganapati, 2020).[1] Uniform confidence bands (UCBs) are very helpful for inferring the true shape, slope, or curvature of h_0, as they graphically convey sampling uncertainty about the estimated structural function and its derivatives.

In applications involving policy counterfactuals, researchers care about estimating and constructing UCBs for h_0 or its derivatives. For instance, Adao et al. (2020, AAG hereafter) derive (1) via a semiparametric gravity equation for the intensive margin of firm exports in a monopolistic competition model based on Melitz (2003). In that context, the derivative of h_0 is the elasticity of the intensive margin of firm-level exports to changes in bilateral trade costs. Moreover, Compiani (2022) performs policy experiments using nonparametric estimates of price elasticities in differentiated product demand models.

As is the case for almost all nonparametric and machine learning (ML) methods, researchers must choose tuning parameters—such as bandwidths, sieve dimensions, or penalty parameters—when estimating or performing inference on h_0 and its derivatives. Poor choice of tuning parameters can lead to estimators that converge unnecessarily slowly and confidence bands with poor coverage. But "good" choices of tuning parameters typically require knowledge of key model regularities, such as the smoothness of h_0 and the strength of the instruments, which are unknown ex ante. It is therefore important to have data-driven methods that *adapt* to unknown model regularities and yield estimators and confidence bands with desirable properties. Data-driven methods for choosing tuning parameters also help to improve the transparency of nonparametric and ML methods,

[1] Other applications include causal inference (Miao, Geng, and Tchetgen Tchetgen, 2018) and reinforcement learning (Chen and Qi, 2022; Chen, Xu, Gulcehre, Paine, Gretton, De Freitas, and Doucet, 2022). Model (1) also nests nonparametric regression when $W = X$, in which case h_0 is the conditional mean of Y given X.

removing a degree of freedom with which the researcher can manipulate results. Unfortunately, popular methods for choosing tuning parameters for nonparametric regression, such as standard cross validation, may not be valid in models with endogeneity—see Section 2.2.

In this paper, we propose simple, data-driven procedures for choosing tuning parameters for estimating and constructing UCBs for h_0 and its derivatives. Our methods are developed for the popular class of sieve nonparametric IV estimators.[2] That is, h_0 is approximated by a linear combination of several basis functions (e.g., B-splines), with the coefficients estimated by Two Stage Least Squares (TSLS) regression of Y on the basis functions of X, using functions of W as instruments (see Section 2.1 for a detailed description). The key tuning parameter to be chosen by a researcher is the number of basis functions, say J, used to approximate h_0. If J is too small, then estimators may be badly biased and UCBs may under-cover. But if J is too large, estimators may be very noisy and UCBs may be uninformatively wide. Before precisely stating our theoretical results in Section 4, we describe our methods and their practical importance.

Our Methods and the Practical Implications. Our first contribution is a data-driven choice of sieve dimension, which we denote by $\tilde{J}$. This choice is simple to compute. Under suitable regularity conditions, we show that sieve estimators implemented with $\tilde{J}$, which we denote $\hat{h}_{\tilde{J}}$, converge at the fastest possible (i.e., minimax) rate in sup-norm.[3] That is, the maximum error over the support of X, namely

$$\sup_x |\hat{h}_{\tilde{J}}(x) - h_0(x)|,$$

vanishes as fast as possible—among *all* estimators of h_0—as the sample size increases, uniformly over a class of data-generating processes (DGPs), for both nonparametric IV and nonparametric regression models. Formally, we refer to $\tilde{J}$ as *sup-norm rate-adaptive*: it adapts to features of the DGP that are unknown ex ante, such as the smoothness of h_0 and strength of the instruments, so that the resulting estimator $\hat{h}_{\tilde{J}}$ converges as fast as possible in sup-norm. We further show that the same data-driven choice $\tilde{J}$ is sup-norm

[2]See Ai and Chen (2003), Newey and Powell (2003), Blundell et al. (2007), and Horowitz (2011).

[3]We focus on the sup-norm rather than L^2 norm (i.e., mean-square error) primarily because our objective is to construct UCBs for h_0 and its derivatives. The sup-norm is essential for this purpose, as we require the entire function (or its derivatives) to lie inside the bands with desired coverage probability. The sup-norm also provides a stronger, more informative sense in which the estimator is converging as it measures the maximal, rather than average, error over the support of X.

rate-adaptive for estimating *derivatives* of h_0 as well.[4] Hence, $\tilde{J}$ should be very useful for researchers interested in estimating elasticities or other marginal effects. We illustrate this usefulness in our empirical application revisiting AAG, where we use $\tilde{J}$ to estimate the elasticity of the intensive margin of firm-level exports from aggregate bilateral trade data. We also demonstrate the good performance of $\tilde{J}$ across a variety of simulation designs for both nonparametric IV estimation and nonparametric regression.

Our second main contribution is a data-driven approach to constructing UCBs for h_0 and its derivatives. The term "uniform" indicates that the entire function lies within the bands with desired asymptotic coverage probability. The UCBs for h_0 and its derivatives are also simple to compute and have strong theoretical justification. They are *honest* in the sense that they guarantee coverage for h_0 and its derivatives uniformly over a generic class of DGPs, and *adaptive* in the sense that they contract at, or within a logarithmic factor of, the minimax rate. As such, they provide efficiency improvements relative to UCBs based on the usual approach of undersmoothing, in which a sub-optimally large J is chosen in the hope that bias is negligible relative to sampling variation. Of course, in empirical work, a researcher does not know the true function, and therefore doesn't know which J is truly large enough that sampling uncertainty dominates bias.

Our UCBs for h_0 and its derivatives are useful for inferring the true shape of the structural function and its derivatives. They complement existing approaches for testing shape restrictions, as they allow the researcher to read off the shape of the function without imposing a specific null (e.g. monotone increasing) a priori. In our empirical application to AAG we construct UCBs for the elasticity of the intensive margin of firm exports. As emphasized by AAG, this is an important, policy-relevant function yet its shape is not restricted by theory in a nonparametric setting. Our UCBs exclude constant functions and downwards-sloping functions. Hence, they provide evidence against the Pareto specification for unobserved firm productivity used by Chaney (2008), which leads to a constant elasticity, as well as other parameterizations used, e.g., by Eaton, Kortum, and Kramarz (2011), Head, Mayer, and Thoenig (2014), and Melitz and Redding (2015), for which the elasticity is downwards-sloping. Empirically-calibrated simulation studies based on the models of Chaney (2008) and Head et al. (2014) demonstrate valid coverage of our UCBs for h_0 and its derivatives and efficiency improvements relative to undersmoothing.

Related Literature and our Theoretical Contributions. Early work on nonparametric IV estimation includes Newey and Powell (2003), Hall and Horowitz (2005), Blun-

[4]This is in contrast to kernel estimation, in which different bandwidths must be used for rate-adaptive estimation of a function and its derivatives.

dell et al. (2007), Darolles, Fan, Florens, and Renault (2011), Horowitz (2011) and others.

We complement prior work by Horowitz (2014) for near-adaptive estimation of h_0 in L^2 norm, Breunig and Johannes (2016) for near-adaptive estimation of linear functionals of h_0, and Breunig and Chen (2021) for adaptive estimation of quadratic functionals of h_0. Our procedure builds on the bootstrap-based implementation of Lepski's method of Chernozhukov, Chetverikov, and Kato (2014) for kernel density estimation and Spokoiny and Willrich (2019) for linear regression with Gaussian errors. But our procedure does not follow easily from theirs due to several challenges present in the conditional moment restriction (1), in which h_0 is identified by $\mathbb{E}[Y|W] = \mathbb{E}[h_0(X)|W]$ (a.s.). The degree of difficulty of inverting $\mathbb{E}[h_0(X)|W]$ to recover h_0 is a nonparametric notion of instrument strength and plays an important role in determining minimax rates for estimators of h_0 and its derivatives.[5] While adaptive procedures for nonparametric density estimation or regression deal only with unknown smoothness of the estimand, our procedures must also deal with the unknown degree of difficulty of the inversion problem. The literature has typically classified the difficulty of the inversion problem into "mild" and "severe" regimes. Minimax rates in the mild regime are achieved by a choice of sieve dimension that balances bias and sampling uncertainty, much like standard nonparametric problems. But minimax rates in the severe regime are obtained by a bias-dominating choice of sieve dimension. Our procedure for data-driven choice of sieve dimension delivers the minimax sup-norm rate for h_0 and its derivatives across the whole spectrum of models, from nonparametric regression to nonparametric IV models in the severe regime.

Our procedure improves significantly on and supersedes a modified Lepski procedure from Section 3 of Chen and Christensen (2015a) on sup-norm rate-adaptive estimation of (1). Ours uses a multiplier bootstrap to avoid selection of several constants and performs much better in practice. Moreover, our rate-adaptivity guarantees encompass nonparametric regression and nonparametric IV in both mild and severe regimes.

Recent work on (non data-driven) UCBs for h_0 and functionals thereof via undersmoothing includes Horowitz and Lee (2012), Chen and Christensen (2018) and Babii (2020). Our UCBs build on prior work on honest, adaptive UCBs for nonparametric density estimation (Giné and Nickl, 2010; Chernozhukov et al., 2014) and Gaussian white noise models (Bull, 2012; Giné and Nickl, 2016). But none of these works allows for nonparametric models with endogeneity, and our procedures do not follow easily from these existing methods due to the above-mentioned challenges present in model (1). Our

[5] See Hall and Horowitz (2005), Chen and Reiss (2011), and Chen and Christensen (2018) for minimax rates for nonparametric IV estimation. When the conditional density of X given W is continuous, these rates are slower than the corresponding rates for nonparametric regression.

UCBs for h_0 and its derivatives apply to nonparametric regression with non-Gaussian, heteroskedastic errors as a special case, which appears to be a new contribution.

Finally, our work also compliments several recent papers on (non data-driven) estimation and inference for nonparametric IV models with shape constraints; see for example Blundell et al. (2017), Chetverikov and Wilhelm (2017), Freyberger and Reeves (2019) and Chernozhukov, Newey, and Santos (2023). These works all assume a deterministic sequence of tuning parameters satisfying regularity conditions that depend on unknown model features such as the smoothness of h_0 and instrument strength. An exception is Breunig and Chen (2020) who study L^2 rate-adaptive testing of a *specific* null hypothesis (e.g., monotone increasing, or a parametric functional form). Our approach is conceptually different from theirs: our UCBs graphically convey sampling uncertainty about an estimate of h_0 and its derivatives. Hence, our UCBs are very useful for inferring the true shape of h_0 in situations—such as our trade application—where there are no specific prior shape restrictions suggested by economic theory.

Outline. Section 2 introduces our methods. Section 3 presents the application to international trade. Section 4 contains the main theoretical results. Section 5 provides additional simulation results for difficult designs. Section 6 presents extensions to additive and partially linear models, and Section 7 concludes. Appendix A presents a simplified version of our procedures for nonparametric regression. Appendix B provides additional details for the trade application and simulations. In the online supplement, Appendix C presents additional simulations to an empirically calibrated Engel curve design, Appendix D gives details on basis functions and nonparametric function classes, and Appendix E contains technical results and proofs.

Notation. Let $\mathcal{X}$ be the support of X, d the dimension of X, and L^2_X and L^2_W the space of functions of X and W with finite second moments. Let $\|h\|_\infty := \sup_{x\in\mathcal{X}} |h(x)|$ be the sup-norm of $h : \mathcal{X} \to \mathbb{R}$. Let $\mathbb{N}$ be the set of integers and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ the non-negative integers. Let $\lceil a \rceil = \min\{n \in \mathbb{N} : n \geq a\}$ and $\lfloor a \rfloor = \max\{n \in \mathbb{N}_0 : n < a\}$. For a multi-index $a = (a_1, ..., a_d) \in (\mathbb{N}_0)^d$ with order $|a| = \sum_{i=1}^d a_i$, the a-derivative of h is defined as

$$\partial^a h(x) = \frac{\partial^{|a|} h(x)}{\partial^{a_1} x_1 \dots \partial^{a_d} x_d}.$$

Let A^- denote the generalized (or Moore–Penrose) inverse of a matrix A and $A^{-1/2}$ the inverse of the positive-definite square root of A.

2 Procedures

We begin in Section 2.1 by briefly reviewing sieve nonparametric IV estimation and UCBs with a deterministic sieve dimension. Section 2.2 explains why standard cross validation for regression fails in models with endogeneity. Section 2.3 presents our data-driven choice of sieve dimension and Section 2.4 presents our data-driven UCBs. These methods extend naturally to partially linear and partially additive models (see Section 6). Both procedures apply to nonparametric regression as well (see Appendix A).

2.1 Review: Estimators and UCBs with a Deterministic J

Estimators. Consider approximating h_0 by a linear combination of J basis functions:

$$h_0(x) \approx (\psi^J(x))'c_J\,, \tag{2}$$

where $\psi^J(x) = (\psi_{J1}(x), \ldots, \psi_{JJ}(x))'$ is a vector of basis functions and $c_J = (c_{J1}, \ldots, c_{JJ})'$ is a vector of coefficients. Combining (1) and (2), we obtain

$$Y = (\psi^J(X))'c_J + \text{bias}_J + u\,, \quad \mathbb{E}[u|W] = 0\,,$$

where $u = Y - h_0(X)$ and $\text{bias}_J = h_0(X) - (\psi^J(X))'c_J$. Provided the bias term is "small" relative to u in an appropriate sense, we have an approximate linear IV model where $\psi^J(X)$ is a $J \times 1$ vector of "endogenous variables" and c_J is a vector of unknown "parameters". One can then estimate c_J using TSLS or GMM using a $K \times 1$ vector of basis functions $b^K(W) = (b_{K1}(W), \ldots, b_{KK}(W))'$ of W as instruments. Evidently, $K \geq J$ is necessary to estimate c_J.

Given data $(X_i, Y_i, W_i)_{i=1}^n$, the TSLS estimator of c_J is simply

$$\hat{c}_J = (\mathbf{\Psi}_J'\mathbf{P}_K\mathbf{\Psi}_J)^-\,\mathbf{\Psi}_J'\mathbf{P}_K\mathbf{Y}\,,$$

where $\mathbf{\Psi}_J = (\psi^J(X_1), \ldots, \psi^J(X_n))'$ and $\mathbf{B}_K = (b^K(W_1), \ldots, b^K(W_n))'$ are $n \times J$ and $n \times K$ matrices, $\mathbf{P}_K = \mathbf{B}_K(\mathbf{B}_K'\mathbf{B}_K)^-\mathbf{B}_K'$ is the projection matrix onto the instrument space, and $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is a $n \times 1$ vector. Estimators of h_0 and its derivative $\partial^a h_0$ are given by

$$\hat{h}_J(x) = (\psi^J(x))'\hat{c}_J\,, \quad \text{and} \quad \partial^a\hat{h}_J(x) = (\partial^a\psi^J(x))'\hat{c}_J\,,$$

where $\partial^a\psi^J(x) = (\partial^a\psi_{J1}(x), \ldots, \partial^a\psi_{JJ}(x))'$.

Sieve Bases. Many linear sieves, such as polynomial splines, B-splines, wavelets, Fourier series, and various polynomials, can be used as the instrument basis $\{b_{Kk}\}_{k=1}^K$. However, only B-splines and Cohen–Daubechies–Vial (CDV) wavelet bases for $\{\psi_{Jj}\}_{j=1}^J$ have been shown to achieve the optimal minimax sup-norm rates under a suitable choice of J (Chen and Christensen, 2018).[6] As our objective is to have estimators that converge as fast as possible in sup-norm—which is essential for constructing UCBs that are as narrow and informative as possible—we restrict attention to B-splines and CDV wavelets for $\{\psi_{Jj}\}_{j=1}^J$ in our theory that follows. Moreover, since B-splines are easy to compute, much less collinear than polynomials and polynomial splines, and available in standard software packages, we confine our presentation to B-spline bases for both $\{\psi_{Jj}\}_{j=1}^J$ and $\{b_{Kk}\}_{k=1}^K$ in the main text.

Key tuning parameter J. Based on simulations and theoretical studies in Blundell et al. (2007), Chen and Christensen (2018) and others, the performance of the sieve TSLS estimator for h_0 is sensitive to the choice of J and not sensitive to K as long as $K \geq J$. We introduce a data-driven method for choosing J in Section 2.3. The choice of K is pinned down by J in our procedure, so we write $K(J) \geq J$, $b^{K(J)}(W)$, $\mathbf{B}_{K(J)}$ and $\mathbf{P}_{K(J)}$ in what follows. Let $\mathbf{M}_J = (\mathbf{\Psi}_J' \mathbf{P}_{K(J)} \mathbf{\Psi}_J)^- \mathbf{\Psi}_J' \mathbf{P}_{K(J)}$ be a $J \times n$ matrix. We can equivalently write

$$\hat{h}_J(x) = (\psi^J(x))' \mathbf{M}_J \mathbf{Y}\,, \qquad \partial^a \hat{h}_J(x) = (\partial^a \psi^J(x))' \mathbf{M}_J \mathbf{Y}\,. \tag{3}$$

"Undersmoothed" UCBs. We now review the usual approach of constructing "undersmoothed" UCBs for h_0 and its derivatives based on a deterministic J. Let $\hat{\mathbf{u}}_J = (\hat{u}_{1,J}, \ldots, \hat{u}_{n,J})'$ denote the $n \times 1$ vector of residuals whose ith element is $\hat{u}_{i,J} = Y_i - \hat{h}_J(X_i)$. Then $\hat{h}_J(x) - h_0(x)$ and $\partial^a \hat{h}_J(x) - \partial^a h_0(x)$ can be estimated by

$$D_J(x) = (\psi^J(x))' \mathbf{M}_J \hat{\mathbf{u}}_J\,, \qquad D_J^a(x) = (\partial^a \psi^J(x))' \mathbf{M}_J \hat{\mathbf{u}}_J\,, \tag{4}$$

and their variances can be estimated by

$$\hat{\sigma}_J^2(x) = (\psi^J(x))' \mathbf{M}_J \widehat{\mathbf{U}}_{J,J} \mathbf{M}_J' \psi^J(x), \qquad \hat{\sigma}_J^{a2}(x) = (\partial^a \psi^J(x))' \mathbf{M}_J \widehat{\mathbf{U}}_{J,J} \mathbf{M}_J' (\partial^a \psi^J(x)) \tag{5}$$

where $\widehat{\mathbf{U}}_{J,J}$ is a $n \times n$ diagonal matrix whose ith diagonal entry is $\hat{u}_{i,J}\hat{u}_{i,J}$.

[6] Bases for h_0 must have bounded Lebesgue constant to attain the minimax sup-norm rate for nonparametric regression (see, e.g., Belloni, Chernozhukov, Chetverikov, and Kato (2015) and Chen and Christensen (2015b)). B-splines and CDV wavelets have this property. Bases without this property, such as polynomials and Fourier series, cannot attain the minimax sup-norm rate and hence cannot lead to sup-norm rate-adaptive estimators or UCBs.

Let $\hat{\mathbf{u}}_J^* = (\hat{u}_{1,J}\varpi_1, \ldots, \hat{u}_{n,J}\varpi_n)'$ denote a multiplier bootstrap version of $\hat{\mathbf{u}}_J$, where $(\varpi_i)_{i=1}^n$ are IID $N(0,1)$ draws independent of the data. Then

$$D_J^*(x) = (\psi^J(x))'\mathbf{M}_J\hat{\mathbf{u}}_J^* \ , \qquad D_J^{a*}(x) = (\partial^a\psi^J(x))'\mathbf{M}_J\hat{\mathbf{u}}_J^* \tag{6}$$

are bootstrap versions of $D_J(x)$ and $D_J^a(x)$. For each independent draw of $(\varpi_i)_{i=1}^n$, compute the sup t-statistics:

$$\sup_{x\in\mathcal{X}}\left|\frac{D_J^*(x)}{\hat{\sigma}_J(x)}\right| \ , \qquad \sup_{x\in\mathcal{X}}\left|\frac{D_J^{a*}(x)}{\hat{\sigma}_J^a(x)}\right| . \tag{7}$$

Let $z_{1-\alpha,J}^*$ and $z_{1-\alpha,J}^{a*}$ denote the $(1-\alpha)$ quantile of these sup statistics across a large number (say 1000) independent draws of $(\varpi_i)_{i=1}^n$. Chen and Christensen (2018) construct $100(1-\alpha)\%$ UCBs for h_0 and $\partial^a h_0$ as follows:

$$C_{n,J}(x) = \left[\hat{h}_J(x) - z_{1-\alpha,J}^*\hat{\sigma}_J(x),\ \hat{h}_J(x) + z_{1-\alpha,J}^*\hat{\sigma}_J(x)\right],$$

$$C_{n,J}^a(x) = \left[\partial^a\hat{h}_J(x) - z_{1-\alpha,J}^{a*}\hat{\sigma}_J^a(x),\ \partial^a\hat{h}_J(x) + z_{1-\alpha,J}^{a*}\hat{\sigma}_J^a(x)\right].$$

The above UCBs are theoretically justified provided J increases faster than the oracle J_0 (the optimal sieve dimension for estimating h_0 or its derivatives in sup-norm), so that the bias is of smaller order than sampling uncertainty. Unfortunately, J_0 is unknown in practice since it depends on the unknown smoothness of h_0 and other unknown model regularities of (1). This motivates us to propose the new data-driven UCBs in Section 2.4.

2.2 Problems with Standard Cross Validation

We briefly explain why the usual approach of cross validation (CV) for regression is not a valid method for choosing J in models with endogeneity. Consider the standard CV criterion

$$\mathrm{CV}(J) = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{h}_{-i,J}(X_i))^2, \tag{8}$$

where n is the sample size and $\hat{h}_{-i,J}$ denotes version of $\hat{h}_J$ computed from a sub-sample that excludes the ith observation. Let $u_i = Y_i - h_0(X_i)$. We may then expand (8) as

$$\mathrm{CV}(J) = \frac{1}{n}\sum_{i=1}^n (h_0(X_i) - \hat{h}_{-i,J}(X_i))^2 + \frac{1}{n}\sum_{i=1}^n u_i^2 + \frac{2}{n}\sum_{i=1}^n u_i(h_0(X_i) - \hat{h}_{-i,J}(X_i)).$$

The first term in the expansion is an estimate of the MSE $\mathbb{E}[(h_0(X) - \hat{h}_J(X))^2]$ of $\hat{h}_J$ and the second term is independent of J. The third term is an estimate of $\mathbb{E}[u(h_0(X) - \hat{h}_J(X))]$. This term is asymptotically negligible without endogeneity (i.e., when $\mathbb{E}[u|X] = 0$) as is the case for nonparametric regression, making $\text{CV}(J)$ a suitable sample analogue of the mean-square error of $\hat{h}_J$ in that case (see, e.g., Li (1987)). But in models with endogeneity (i.e., when $\mathbb{E}[u|X] \neq 0$), there is no guarantee that $\mathbb{E}[u(h_0(X) - \hat{h}_J(X))] = 0$ and so this third term—which depends on J—may be non-negligible even asymptotically. If so, cross validation gives a biased estimate of the MSE of $\hat{h}_J$ and is therefore not a meaningful criterion by which to choose J in models with endogeneity. Indeed, a cross-validated choice of J may not even lead to a consistent estimator of h_0 in model (1).

In addition, even for nonparametric regression, the J chosen by CV balances bias and sampling uncertainty in L^2 norm. Such as choice is not optimal for estimation of h_0 and its derivatives in sup-norm, nor is it sutiable for adaptive UCBs for h_0 and its derivatives.

2.3 Procedure 1: Data-driven Choice of Sieve Dimension

We now present our data-driven choice $\tilde{J}$ of sieve dimension using B-spline bases. B-splines are characterized by their order r. In the simulations and empirical application, we use a cubic B-spline ($r = 4$) for $\{\psi_{Jj}\}_{j=1}^J$ and a quartic B-spline ($r = 5$) for $\{b_{Kk}\}_{k=1}^K$.[7]

Let $\mathcal{T} = \{J = (2^l + r - 1)^d : l \in \mathbb{N}_0\}$ denote a dyadic grid of candidate values of J, where the integer r is the order of the B-spline basis for $\{\psi_{Jj}\}_{j=1}^J$ (i.e., each ψ_{Jj} is a piecewise polynomial of degree $r - 1$). For example, $\mathcal{T} = \{J = 2^l + 3 : l \in \mathbb{N}_0\} \equiv \{4, 5, 7, 11, 19, 35, \ldots\}$ for a scalar X ($d = 1$) and cubic B-splines ($r = 4$).[8] The index l is the *resolution level*. We construct $\{b_{Kk}\}_{k=1}^K$ similarly, using B-splines of order $(r + 1)$ because the reduced form is smoother than h_0. Given the resolution level l for the basis for X, the resolution level for the basis for W is $l_w = \lceil (l+q)d/d_w \rceil$ for some $q \in \mathbb{N}_0$ where d_w is the dimension of W. Linking l_w to l in this manner defines a mapping $K(J)$ that satisfies $\lim_{J\to\infty} K(J)/J = c \in [1, \infty)$. We recommend taking q as the second- or third-smallest value for which $K(J) \geq J$ holds for all J (i.e., $q = 1$ or $q = 2$ if both X and W are of the same dimension). We advise against choosing q any larger, as the number of basis functions increases exponentially in the resolution level. Let $J^+ = \min\{j \in \mathcal{T} : j > J\}$ be the smallest sieve dimension in $\mathcal{T}$ exceeding J.

[7] In the first submitted version we also used a quadratic B-spline ($r = 3$) for $\{\psi_{Jj}\}_{j=1}^J$. In additional simulations we obtained very similar results with a Fourier basis for $\{b_{Kk}\}_{k=1}^K$.

[8] Letting J vary over $\mathcal{T}$ ensures there is enough separation that we can accurately compare the bias and variance of estimators with different $J \in \mathcal{T}$. This helps improve the numerical stability of the method, coherent with implementations of Lepski's method in other nonparametric contexts.

For $J, J_2 \in \mathcal{T}$ with $J_2 > J$, the contrast $D_J(x) - D_{J_2}(x)$ is an estimate of $\hat{h}_J(x) - \hat{h}_{J_2}(x)$, whose variance can be estimated by

$$\hat{\sigma}^2_{J,J_2}(x) := \hat{\sigma}^2_J(x) + \hat{\sigma}^2_{J_2}(x) - 2\tilde{\sigma}_{J,J_2}(x), \quad \tilde{\sigma}_{J,J_2}(x) = (\psi^J(x))'\mathbf{M}_J\widehat{\mathbf{U}}_{J,J_2}\mathbf{M}'_{J_2}\psi^{J_2}(x), \tag{9}$$

where $\hat{\sigma}^2_J(x)$ is defined in (5) and $\widehat{\mathbf{U}}_{J,J_2}$ is a $n \times n$ diagonal matrix whose ith diagonal entry is $\hat{u}_{i,J}\hat{u}_{i,J_2}$. Moreover, the multiplier bootstrap version of $D_J(x) - D_{J_2}(x)$ is

$$D^*_J(x) - D^*_{J_2}(x) = (\psi^J(x))'\mathbf{M}_J\hat{\mathbf{u}}^*_J - (\psi^{J_2}(x))'\mathbf{M}_{J_2}\hat{\mathbf{u}}^*_{J_2}.$$

Finally let $\hat{s}_J$ be the smallest singular value of $(\mathbf{B}'_{K(J)}\mathbf{B}_{K(J)})^{-1/2}(\mathbf{B}'_{K(J)}\mathbf{\Psi}_J)(\mathbf{\Psi}'_J\mathbf{\Psi}_J)^{-1/2}$.

Procedure 1: Data-driven Choice of Sieve Dimension

1. Compute

$$\hat{J}_{\max} = \min\left\{J \in \mathcal{T} : J\sqrt{\log J}\hat{s}_J^{-1} \leq 10\sqrt{n} < J^+\sqrt{\log J^+}\hat{s}_{J^+}^{-1}\right\} \tag{10}$$

$$\hat{\mathcal{J}} = \left\{J \in \mathcal{T} : 0.1(\log \hat{J}_{\max})^2 \leq J \leq \hat{J}_{\max}\right\}. \tag{11}$$

2. Let $\hat{\alpha} = \min\{0.5, (\log(\hat{J}_{\max})/\hat{J}_{\max})^{1/2}\}$. For each draw of $(\varpi_i)_{i=1}^n$, compute

$$\sup_{\{(x,J,J_2)\in\mathcal{X}\times\hat{\mathcal{J}}\times\hat{\mathcal{J}}:J_2>J\}}\left|\frac{D^*_J(x) - D^*_{J_2}(x)}{\hat{\sigma}_{J,J_2}(x)}\right|. \tag{12}$$

   Let $\theta^*_{1-\hat{\alpha}}$ denote the $(1-\hat{\alpha})$ quantile of (12) across independent draws of $(\varpi_i)_{i=1}^n$.

3. Let $\hat{J}_n = \max\{J \in \hat{\mathcal{J}} : J < \hat{J}_{\max}\}$ and

$$\hat{J} = \min\left\{J \in \hat{\mathcal{J}} : \sup_{(x,J_2)\in\mathcal{X}\times\hat{\mathcal{J}}:J_2>J}\left|\frac{\hat{h}_J(x) - \hat{h}_{J_2}(x)}{\hat{\sigma}_{J,J_2}(x)}\right| \leq 1.1\theta^*_{1-\hat{\alpha}}\right\}. \tag{13}$$

   The data-driven choice of sieve dimension is

$$\tilde{J} = \min\{\hat{J}, \hat{J}_n\}. \tag{14}$$

Remark 2.1 In practice, the supremums over x in Steps 2 and 3 can be replaced by the maximum over a fine grid of x values as the functions are continuous in x. We have used 1000 draws of $(\varpi_i)_{i=1}^n$ in our empirical and simulation studies. Note the $(\varpi_i)_{i=1}^n$ are held

fixed when computing the supremum over (x, J, J_2) for each draw. Our theory allows for constants other than 10 and 0.1 in Step 1 as long as they ensure $\hat{\mathcal{J}}$ contains several values of J to search over. Our theory also allows for any constant larger than 1 in Step 3; the value 1.1 performed well in simulations and is used in other implementations of Lepski's method (see, e.g., Chernozhukov et al. (2014)).

We present the theoretical results on the adaptivity of $\tilde{J}$ in Section 4.2.

2.4 Procedure 2: Data-driven UCBs

Let $\underline{p} > d/2$ denote the minimal degree of smoothness assumed for h_0. For instance, if X is scalar and h_0 is Lipschitz, then one could take $\underline{p} = 1$ even through the true smoothness of h_0 is unknown. Let $\hat{A} = \log\log \tilde{J}$ and

$$\hat{\mathcal{J}}_- = \begin{cases} \{J \in \hat{\mathcal{J}} : J < \hat{J}_n\} & \text{if } \tilde{J} = \hat{J}, \\ \hat{\mathcal{J}} & \text{if } \tilde{J} = \hat{J}_n. \end{cases}$$

Procedure 2: Data-driven UCBs for h_0

4. For each $(\varpi_i)_{i=1}^n$, compute

$$\sup_{(x,J)\in\mathcal{X}\times\hat{\mathcal{J}}_-} \left| \frac{D_J^*(x)}{\hat{\sigma}_J(x)} \right|. \tag{15}$$

   Let $z_{1-\alpha}^*$ denote the $(1-\alpha)$ quantile of (15) across independent draws of $(\varpi_i)_{i=1}^n$.

5. Construct the $100(1-\alpha)\%$ UCB

$$C_n(x) = \left[\hat{h}_{\tilde{J}}(x) - \mathrm{cv}^*(x)\, \hat{\sigma}_{\tilde{J}}(x),\ \hat{h}_{\tilde{J}}(x) + \mathrm{cv}^*(x)\, \hat{\sigma}_{\tilde{J}}(x) \right], \tag{16}$$

   where

$$\mathrm{cv}^*(x) = \begin{cases} z_{1-\alpha}^* + \hat{A}\theta_{1-\hat{\alpha}}^* & \text{if } \tilde{J} = \hat{J}, \\ z_{1-\alpha}^* + \hat{A}\max\{\theta_{1-\hat{\alpha}}^*\,,\ \tilde{J}^{-\underline{p}/d}/\hat{\sigma}_{\tilde{J}}(x)\} & \text{if } \tilde{J} = \hat{J}_n. \end{cases} \tag{17}$$

Procedure 2′: Data-driven UCBs for $\partial^a h_0$ $(0 < |a| < \underline{p})$

4′. For each $(\varpi_i)_{i=1}^n$, compute

$$\sup_{(x,J)\in\mathcal{X}\times\hat{\mathcal{J}}_-} \left| \frac{D_J^{a*}(x)}{\hat{\sigma}_J^a(x)} \right|. \tag{18}$$

   Let $z_{1-\alpha}^{a*}$ denote the $(1-\alpha)$ quantile of (18) across independent draws of $(\varpi_i)_{i=1}^n$.

5′. Construct the $100(1-\alpha)\%$ UCB

$$C_n^a(x) = \left[\partial^a \hat{h}_{\tilde{J}}(x) - \mathrm{cv}^{a*}(x)\,\hat{\sigma}^a_{\tilde{J}}(x),\ \partial^a \hat{h}_{\tilde{J}}(x) + \mathrm{cv}^{a*}(x)\,\hat{\sigma}^a_{\tilde{J}}(x)\right], \tag{19}$$

where

$$\mathrm{cv}^{a*}(x) = \begin{cases} z^{a*}_{1-\alpha} + \hat{A}\theta^*_{1-\hat{\alpha}} & \text{if } \tilde{J} = \hat{J}, \\ z^{a*}_{1-\alpha} + \hat{A}\max\{\theta^*_{1-\hat{\alpha}}, \tilde{J}^{(|a|-\underline{p})/d}/\hat{\sigma}^a_{\tilde{J}}(x)\} & \text{if } \tilde{J} = \hat{J}_n. \end{cases} \tag{20}$$

Remark 2.2 Procedures 1 and 2 require choosing the B-spline order r and Procedure 2 requires specifying the minimal degree of smoothness $\underline{p}$. For sup-norm estimation and UCBs for first derivatives one can take $r \geq 3$ and $\underline{p} \geq 1$; for second derivatives and cross elasticities one can take $r \geq 4$ and $\underline{p} \geq 2$.

Remark 2.3 We establish that $\tilde{J} = \hat{J}$ with probability approaching one (wpa1) in the mild regime; and that $\tilde{J}, \hat{J} \in [c\hat{J}_n, \hat{J}_n]$ wpa1 in the severe regime (for a constant $c \in (0,1)$). Nevertheless, we find $\tilde{J} = \hat{J}$ in the empirical application and in the vast majority (between 99.6% and 100% depending on the design and sample size) of all simulations. In particular, $\tilde{J} = \hat{J}$ across all simulations in the Engel curve design which is in the severe regime (see Appendix C).

Theoretical properties of these UCBs are presented in Sections 4.3 and 4.4. We show that the Procedures 2 and 2′ UCBs are honest and adaptive for models in the mild regime (including nonparametric regression as a special case). For models in the severe regime, we show that the Procedures 2 and 2′ UCBs with critical values corresponding to $\tilde{J} = \hat{J}_n$ have valid (actually conservative) coverage. Nevertheless, the Engel curve simulation in Appendix C shows that the Procedure 2 UCBs still have valid (actually conservative) coverage for a severe regime design.

3 International Trade: Simulations and Application

Adao, Arkolakis, and Ganapati (2020, hereafter AAG) derive semiparametric gravity equations for the extensive and intensive margins of firm exports in a monopolistic competition model of international trade. Importantly, and in sharp contrast with the existing literature (Melitz, 2003; Chaney, 2008; Eaton et al., 2011; Head et al., 2014; Melitz and Redding, 2015), AAG do not impose any parametric assumptions on the distribution of unobserved

firm heterogeneity. The gravity equations identify functions which characterize the elasticities of the extensive and intensive margins of firm-level exports to changes in bilateral trade costs. AAG emphasize the importance of these elasticities for counterfactuals.

In this section, we apply our procedures to estimate and construct UCBs for the intensive margin and its elasticity using AAG's baseline model and data. We also present simulation studies based on empirical calibrations of two workhorse trade models to illustrate the sound performance of our procedures.

3.1 Model and Data

We begin by briefly summarizing the empirical framework of AAG. They use a monopolistic competition model of international trade—see Melitz and Redding (2014) for a review. There are a continuum of firms in each country. Firm ω in country i is characterized by an entry potential $e_{ij}(\omega)$ and a revenue potential $r_{ij}(\omega)$ for selling in country j. Firms draw $e_{ij}(\omega)$ from a distribution $H^e_{ij}(e)$ then $r_{ij}(\omega)$ from a (possibly degenerate) distribution $H^r_{ij}(r|e)$. Firm ω in country i exports to country j if and only if $e_{ij}(\omega)$ exceeds a threshold. The proportion of firms in country i that export to country j is denoted π_{ij}.

The extensive margin is characterized by the inverse distribution of entry potential, i.e., $\epsilon_{ij}(\pi_{ij}) = (H^e_{ij})^{-1}(1-\pi_{ij})$. Assuming homogeneity (so $H^e_{ij} = H^e$ and $\epsilon_{ij} = \epsilon$), AAG's gravity equation for the extensive margin is

$$\log \epsilon(\pi_{ij}) = \log(\bar{f}_{ij}\bar{\tau}_{ij}^{\tilde{\sigma}}) + \delta_i^\epsilon + \zeta_j^\epsilon,$$

where $\bar{\tau}_{ij}$ and $\bar{f}_{ij}$ are variable and fixed trade costs from i to j and δ_i^ϵ and ζ_j^ϵ are exporter and importer fixed effects (FEs). Costs depend linearly on a cost shifter z_{ij}:

$$\begin{aligned}\log \bar{\tau}_{ij} &= \kappa^\tau z_{ij} + \delta_i^\tau + \zeta_j^\tau + \eta_{ij}^\tau,\\ \log \bar{f}_{ij} &= \kappa^f z_{ij} + \delta_i^f + \zeta_j^f + \eta_{ij}^f,\end{aligned}$$

where the idiosyncratic error terms η_{ij}^τ and η_{ij}^f are conditionally mean-zero and independent of z_{ij} and the FEs. This yields the estimating equation

$$\log \epsilon(\pi_{ij}) = (\kappa^f + \tilde{\sigma}\kappa^\tau)z_{ij} + (\delta_i^f + \tilde{\sigma}\delta_i^\tau + \delta_i^\epsilon) + (\zeta_j^f + \tilde{\sigma}\zeta_j^\tau + \zeta_j^\epsilon) + \eta_{ij}^f + \tilde{\sigma}\eta_{ij}^\tau. \tag{21}$$

Note that π_{ij} depends (possibly nonlinearly) on z_{ij} and the error terms η_{ij}^f and η_{ij}^τ.

The intensive margin is characterized by the average revenue potential of exporting

firms:

$$\rho_{ij}(\pi) = \frac{1}{\pi}\int_0^{\pi} \mathbb{E}[r|e = \epsilon_{ij}(v)]\,\mathrm{d}v\,,$$

where the expectation is taken under $H^r_{ij}(r|e)$. Assuming homogeneity (so $H^r_{ij} = H^r$ and $\rho_{ij} = \rho$), AAG's gravity equation for the intensive margin is

$$\log \bar{x}_{ij} - \log \rho(\pi_{ij}) = \log(\bar{\tau}_{ij}^{\tilde{\sigma}}) + \delta_i^{\rho} + \zeta_j^{\rho},$$

where $\bar{x}_{ij}$ are average firm exports and δ_i^{ρ} and ζ_j^{ρ} are FEs. With $\bar{\tau}_{ij}$ as above, AAG obtain

$$\log \bar{x}_{ij} + \tilde{\sigma}\kappa^{\tau} z_{ij} = \log \rho(\pi_{ij}) + (\delta_i^{\rho} - \tilde{\sigma}\delta_i^{\tau}) + (\zeta_j^{\rho} - \tilde{\sigma}\zeta_j^{\tau}) - \tilde{\sigma}\eta_{ij}^{\tau}. \tag{22}$$

More concisely,

$$y_{ij} = \log \tilde{\rho}(\tilde{\pi}_{ij}) + \delta_i + \zeta_j + u_{ij}\,, \tag{23}$$

where $y_{ij} := \log \bar{x}_{ij} + \tilde{\sigma}\kappa^{\tau} z_{ij}$ is the dependent variable,[9] $\tilde{\pi}_{ij} := \log \pi_{ij}$ is the endogenous regressor, $\log \tilde{\rho}(\tilde{\pi}) := \log \rho(e^{\tilde{\pi}})$ is the unknown structural function, $\delta_i := \delta_i^{\rho} - \tilde{\sigma}\delta_i^{\tau}$ and $\zeta_j := \zeta_j^{\rho} - \tilde{\sigma}\zeta_j^{\tau}$ are exporter and importer FEs, and the idiosyncratic error term $u_{ij} := -\tilde{\sigma}\eta_{ij}^{\tau}$ is conditionally mean-zero and independent of the instrumental variable z_{ij}.

Our goal is to use (23) to estimate $\log \tilde{\rho}$ and its derivative, as $\frac{\partial \log \tilde{\rho}(\tilde{\pi})}{\partial \tilde{\pi}} \equiv \frac{\partial \log \rho(\pi)}{\partial \log \pi}$ characterizes the elasticity of the intensive margin of firm-level exports to changes in bilateral trade costs. We use the same data that AAG use for their baseline estimates, which consists of $\bar{x}_{ij}$, z_{ij}, and $\tilde{\pi}_{ij}$ for a sample of 1522 country pairs for the year 2012. We refer the reader to AAG for a detailed description of the data and its construction.

3.2 Implementation

Model (23) differs from model (1) due to the presence of FEs. AAG estimate $\log \tilde{\rho}$ and FEs jointly, using both z_{ij} and exporter and importer country dummies as instruments. As such, they estimate a partially linear model with a large number of linear regressors (due to the country dummies) and, similarly, a large number of instrumental variables.[10] Our methods and theoretical results are not formally developed for such a setting.[11] Therefore, we maintain their assumption that z_{ij} and origin and destination FEs are

[9] AAG construct y_{ij} from data on $\bar{x}_{ij}$ and z_{ij} based on external estimates of $\tilde{\sigma}$ and κ^{τ}.

[10] These comments are based on the November 2020 version of AAG, which is currently under revision. Some of their implementation and findings may differ in future versions.

[11] Our approach extends to partially linear models—see Section 6. But with bilateral trade data the number of dummy variables representing origin and destination FEs is increasing with the sample size n. This "many regressors/many instruments" asymptotic framework falls outside the scope of our analysis.

exogenous, but we further assume that $\mathbb{E}[\log \tilde{\rho}(\tilde{\pi}_{ij})|z_{ij}, \delta_i, \zeta_j] = \mathbb{E}[\log \tilde{\rho}(\tilde{\pi}_{ij})|z_{ij}]$ (a.s.). That is, the intensive margin is conditional mean independent of exporter- and importer-specific factors given cost shifters. Note, however, that we are not imposing that average firm exports are conditional mean independent of exporter- and importer-specific factors. The reduced form for y_{ij} is

$$y_{ij} = g(z_{ij}) + \delta_i + \zeta_j + e_{ij}\,, \tag{24}$$

where $g(z_{ij}) = \mathbb{E}[\log \tilde{\rho}(\tilde{\pi}_{ij})|z_{ij}]$ and $\mathbb{E}[e_{ij}|z_{ij}, \delta_i, \zeta_j] = 0$. We estimate δ_i and ζ_j from (24) by partially linear series regression. That is, we regress y_{ij} on origin and destination dummies and functions $b_{K1}, \ldots, b_{KK}$ of z_{ij} at dimension $K(\hat{J}_{\max})$. We then apply our procedures using $Y_{ij} = y_{ij} - \hat{\delta}_i - \hat{\zeta}_j$ as the dependent variable (Y), $\tilde{\pi}_{ij}$ as the endogenous regressor (X), and z_{ij} as the instrumental variable (W). We present simulations below for models with and without FEs and show that this first-stage estimation of δ_i and ζ_j does not affect the performance of our procedures. Appendix B provides further details on implementation.

3.3 Empirical Results

We implement our procedures using AAG's data. Our data-driven choice of sieve dimension is $\tilde{J} = 4$ for this sample. Figure 1 plots our estimate of $\log \rho$ and the elasticity of the intensive margin, together with their 95% UCBs that are constructed as in displays (16) and (19), respectively. We report results over the interval $[0.1\%, 50\%]$, as in AAG.

UCBs for $\log \rho$ and the elasticity of ρ are both narrow and informative. Figure 1 also plots a linear IV estimate of $\log \rho$ and the corresponding (constant) elasticity estimate.[12] These both lie outside the UCBs for much of the support of π_{ij}. As such, our UCBs for the elasticity provide evidence against the Pareto specification for unobserved firm productivity used, e.g., by Chaney (2008), under which the elasticity of ρ is constant. Whereas Figure 1 of AAG shows that several conventional parameterizations of the distribution of unobserved firm heterogeneity used by Eaton et al. (2011), Head et al. (2014), and Melitz and Redding (2015) all imply a *decreasing* elasticity over $[0.1\%, 50\%]$. By contrast, decreasing elasticities necessarily fall outside our 95% UCBs over $[0.1\%, 50\%]$, as the right-most point of the lower UCB lies above the upper UCB for smaller values of $\tilde{\pi}_{ij}$.

To show that our results are not sensitive to first-stage elimination of fixed effects, we also estimate $\log \rho$ and the FEs jointly, using our data-driven choice $\tilde{J} = 4$ and instrumenting with $b_{K(\tilde{J})1}(z_{ij}), \ldots, b_{K(\tilde{J})K(\tilde{J})}(z_{ij})$ and the origin and destination dummies, and using

[12] For the linear IV estimates, we estimate $\log \rho$ jointly with the FEs as in AAG.

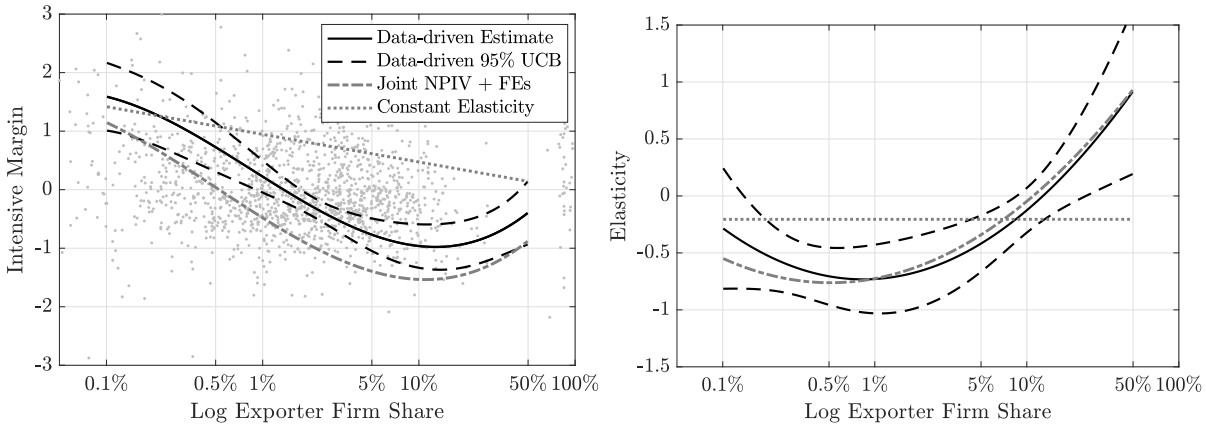


Figure 1: Estimates of the intensive margin $\log \rho$ (left panel) and its elasticity (right panel) using AAG's data set (1522 observations). *Note*: Solid black lines are estimates; dashed black lines are 95% UCBs; dot-dash grey lines are nonparametric estimates with FEs estimated jointly with $\log \rho$ as in AAG; dotted grey lines are linear IV estimates.

y_{ij} as the dependent variable. Estimates using this approach are also shown in Figure 1 (labeled Joint NPIV + FEs). There is a vertical shift in the estimate of $\log \rho$ between the two approaches due to the different treatment of FEs, but the estimated elasticity—which is the focus of AAG—lies entirely within our 95% UCB for the elasticity and is very close to our data-driven elasticity estimate over the whole range $[0.1\%, 50\%]$.

3.4 Simulation Results

We now present simulation studies based on empirical calibrations of two workhorse trade models. The first design is based on Head et al. (2014) who assume a log-normal distribution for latent firm productivity. The second design is based on Chaney (2008) who assumes a Pareto distribution. In the first design the elasticity of ρ is decreasing whereas in the second design $\log \rho(\pi) = \rho \log \pi$ and hence the elasticity is constant. For brevity we only present results for elasticity estimates in the log-normal design here. Additional results for the Pareto design and estimation of $\log \rho$ are deferred to Appendix B.2.

We generate data by first sampling z_{ij} independently with replacement from its empirical distribution. We then generate data on $\tilde{\pi}_{ij}$ and $\bar{x}_{ij}$ by simulating from equations (21) and (22), using the expressions for $\log \epsilon(\pi)$ and $\log \rho(\pi)$ implied by the log-normal assumption—see Appendix B.2. As the empirical application has $n = 1522$, we investigate the performance of our procedures across 1000 samples of size 761, 1522, 3044, and 6088.

Plots for a representative sample of size 1522 are presented in Figure 2(a). We generate the results in Table 1 and Figure 2 by implementing our procedures as in the empirical

application. That is, the dependent variable is $Y_{ij} = y_{ij} - \hat{\delta}_i - \hat{\zeta}_j$, where $\hat{\delta}_i$ and $\hat{\zeta}_j$ are first-stage estimates of the exporter and importer fixed effects. We construct basis functions as in the application; see Appendix B for details. We also compute estimates and confidence bands over the range 0.1% to 50% for π_{ij} as reported in the application.

The first panel in Table 1 presents the average and median (across simulations) of

$$\sup_{\pi \in [0.001, 0.5]} \left| \frac{d\widehat{\log \rho}(\pi)}{d \log \pi} - \frac{d \log \rho(\pi)}{d \log \pi} \right|,$$

which is the maximal error of estimates of the elasticity of ρ for π_{ij} over $[0.1\%, 50\%]$. We compare estimates using $\tilde{J}$ to estimates that use a deterministic choice of sieve dimension, namely $J = 4$, 5, 7, and 11 (these are the first few values of J over which our procedure searches). In each simulation, the maximal error is generally smallest with $J = 4$ or $J = 5$. The average $\tilde{J}$ is between 4.1 and 4.2 depending on the sample size. The maximal error of $\tilde{J}$ is at least half that with $J = 7$, and ten times smaller than with $J = 11$.

Turning to the coverage properties of UCBs for the elasticity, the second panel of Table 1 shows our data-driven UCBs have correct but somewhat conservative coverage. Some conservativeness is to be expected, as our UCBs have uniform coverage guarantees over a class of DGPs. We also present coverage of UCBs based on the usual approach of "undersmoothing" from Section 2.1. These UCBs use a deterministic J and have valid coverage provided J is chosen sufficiently large that bias is negligible relative to sampling uncertainty. Of course, in any empirical application a researcher does not know the true function, and therefore doesn't know which values of J are sufficiently large that sampling uncertainty dominates bias. As can be seen from Table 1, $J = 4$ or $J = 5$ seems too small, and consequently these bands under-cover. Bands with $J = 7$ have coverage closer to nominal coverage, but these bands are more than 70% wider than the data-driven bands. Comparing the UCBs in Figures 2(a) and 2(c), we see the efficiency improvement of our bands relative to undersmoothed bands with $J = 7$, for estimating both ρ and its elasticity.

The fact that our UCBs are based on an optimal choice of J, and therefore contract faster than bands based on undersmoothing, has important practical consequences. Consider the data-driven UCBs for the elasticity of ρ reported in Figure 2(a). These bands do not contain any constant function because the upper limit of the lower band exceeds the lower limit of the upper band. This provides evidence against the Pareto specification of productivity used by Chaney (2008), for which the elasticity of ρ is constant.[13] Note

[13] Table 1 presents the frequency that such a test rejects the constant elasticity specification.

Table 1: Simulation Results for the Elasticity of ρ, Log-normal Design

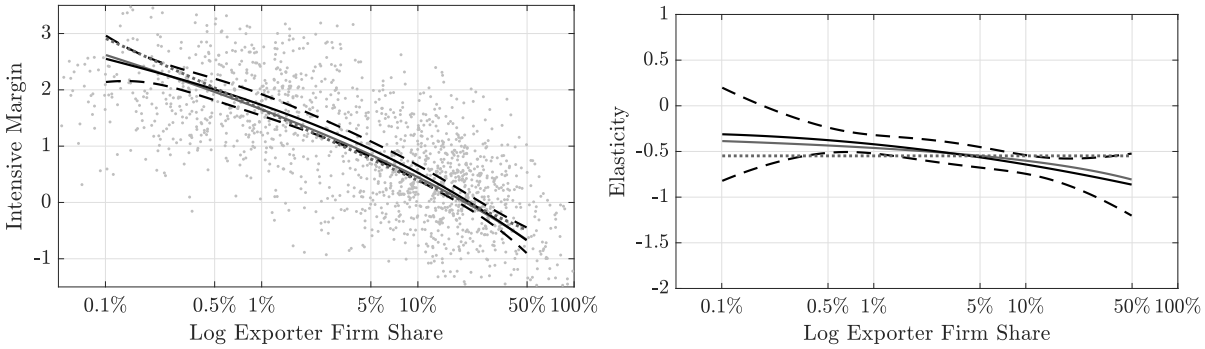
	Data-driven		Deterministic							
			$J=4$		$J=5$		$J=7$		$J=11$	
					Sup-norm Loss					
n	mean	med.	mean	med.	mean	med.	mean	med.	mean	med.
761	0.268	0.187	0.207	0.178	0.314	0.281	0.579	0.472	2.063	1.902
1522	0.184	0.129	0.144	0.125	0.216	0.191	0.382	0.339	1.823	1.650
3044	0.143	0.099	0.106	0.095	0.149	0.139	0.283	0.254	1.562	1.385
6088	0.111	0.071	0.076	0.068	0.105	0.096	0.202	0.185	1.367	1.218
					UCB Coverage					
	90%	95%	90%	95%	90%	95%	90%	95%	90%	95%
761	0.989	0.997	0.861	0.921	0.841	0.911	0.871	0.930	0.906	0.965
1522	0.994	0.997	0.872	0.924	0.857	0.921	0.889	0.936	0.940	0.976
3044	0.993	0.998	0.833	0.899	0.869	0.929	0.899	0.943	0.947	0.979
6088	0.993	0.994	0.800	0.890	0.868	0.936	0.899	0.952	0.949	0.982
	Frequency		95% UCB Relative Width (Deterministic/Data-driven)							
	reject		mean	med.	mean	med.	mean	med.	mean	med.
761	0.088		0.624	0.651	0.922	0.932	1.750	1.568	6.398	6.140
1522	0.344		0.632	0.657	0.906	0.911	1.739	1.599	8.295	8.098
3044	0.822		0.638	0.657	0.888	0.902	1.746	1.625	10.340	10.071
6088	0.959		0.634	0.660	0.865	0.893	1.722	1.690	12.783	12.665

Note: Column "reject" reports the proportion of simulations in which constant functions are excluded from data-driven 95% UCBs for the elasticity.

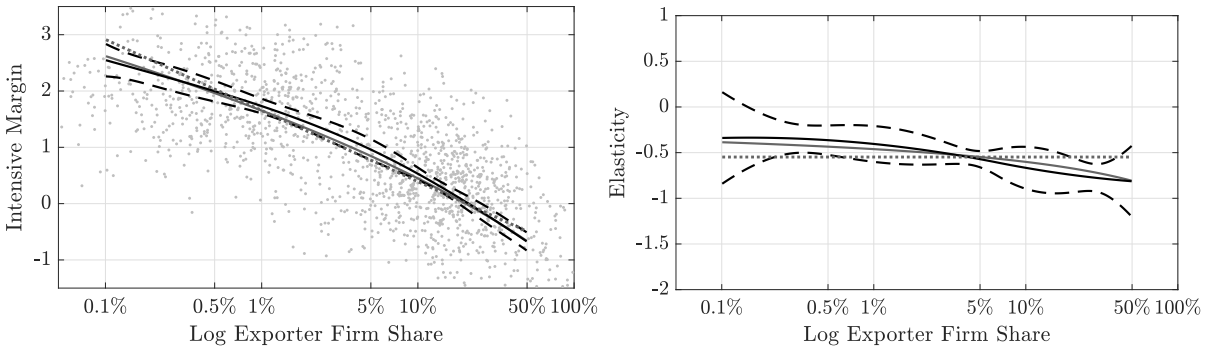
this is despite the fact that our bands tend to be a bit conservative. The undersmoothed bands with $J = 7$ have coverage closer to nominal coverage. But for the sample shown in Figure 2, the undersmoothed bands with $J = 7$ are sufficiently wide that constant functions lie entirely within the bands. Hence, the researcher could not reject a constant elasticity specification on the basis of the undersmoothed bands in this sample. In fact, the undersmoothed bands with $J = 7$ only reject the constant elasticity specification in 15.8% of simulations with 1522 observations whereas the rejection rate for the data-driven bands is 34.4%. This difference in rejection rates illustrates the general phenomenon that undersmoothed bands sacrifice efficiency for coverage. The undersmoothed bands are also quite wiggly, making it difficult to infer the shape of the true elasticity.

We note in closing that our procedures can equally be applied to other IV-based nonparametric analyses in international trade; see, e.g., Adao et al. (2017).

(a) Data-driven Estimates and UCBs



(b) Estimates and UCBs with $J = 5$



(c) Estimates and UCBs with $J = 7$

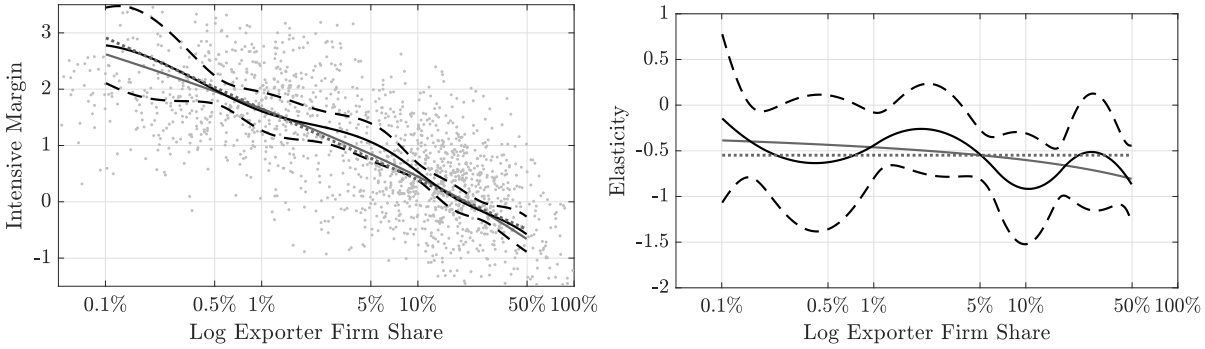


Figure 2: Log-normal design: Plots for a representative sample of size 1522. Left panels correspond to the intensive margin, right panels correspond to its elasticity. *Note:* Solid grey lines are the true curves; solid black lines are estimates; dashed black lines are 95% UCBs; dotted grey lines are linear IV estimates.

4 Theory

We first outline the main regularity conditions in Section 4.1. Section 4.2 shows that $\tilde{J}$ leads to minimax convergence rates for estimators of both h_0 and its derivatives. We then present the main results for UCBs in Sections 4.3 and 4.4.

4.1 Assumptions

We first state and then discuss the assumptions that we impose on the model and sieve space. We require these to hold for some constants $a_f, \underline{c}, \overline{C}, C_T, C_Q, \underline{\sigma}, \overline{\sigma} > 0$ and $\gamma \in (0,1)$. Let $T : L^2_X \to L^2_W$ denote the operator $Th(w) = \mathbb{E}[h(X)|W = w]$. For nonparametric regression we have $W \equiv X$ and so T reduces to the identity.

Assumption 1 *(i) X has support $\mathcal{X} = [0,1]^d$ and its distribution has Lebesgue density f_X which satisfies $a_f^{-1} < f_X(x) < a_f$ on $\mathcal{X}$; (ii) W has support $\mathcal{W} = [0,1]^{d_w}$ and its distribution has Lebesgue density f_W which satisfies $a_f^{-1} < f_W(w) < a_f$ on $\mathcal{W}$; (iii) T is injective.*

Assumption 2 *(i)* $\mathbb{P}\big(\mathbb{E}[u^4|W] \leq \overline{\sigma}^2\big) = 1$*; (ii)* $\mathbb{P}\big(\mathbb{E}[u^2|W] \geq \underline{\sigma}^2\big) = 1$.

Let Ψ_J and B_K be the closed linear subspaces of L^2_X and L^2_W spanned by $\psi_{J1}, \ldots, \psi_{JJ}$ and $b_{K1}, \ldots, b_{KK}$, respectively. Define

$$\tau_J = \sup_{h \in \Psi_J : \|h\|_{L^2_X} \neq 0} \frac{\|h\|_{L^2_X}}{\|Th\|_{L^2_W}},$$

where $\|\cdot\|_{L^2_X}$ and $\|\cdot\|_{L^2_W}$ denote the L^2_X and L^2_W norms. The *sieve measure of ill-posedness* τ_J quantifies the degree of difficulty of inverting Th_0 to recover h_0. As conditional expectations are (weakly) contractive, we have $\tau_J \geq 1$. Large τ_J indicate a more difficult inversion problem. The model (1) is said to be *mildly ill-posed* (or in the *mild* regime) if $\tau_J \asymp J^{\varsigma/d}$ for some $\varsigma \geq 0$ and *severely ill-posed* (or in the *severe* regime) if $\tau_J \asymp \exp(CJ^{\varsigma/d})$ for some $C, \varsigma > 0$, where $d = \dim(X)$. For nonparametric regression models we have $\tau_J = 1$ for all J. Hence, nonparametric regression is a special case of the mild regime with $\varsigma = 0$.

Let $\Pi_J : L^2_X \to \Psi_J$ and $\Pi_{K(J)} : L^2_W \to B_{K(J)}$ denote LS projections onto Ψ_J and $B_{K(J)}$:

$$\Pi_J f = \arg\min_{g \in \Psi_J} \|f - g\|_{L^2_X}, \quad \Pi_{K(J)} f = \arg\min_{g \in B_{K(J)}} \|f - g\|_{L^2_W}.$$

Also let $Q_J : L^2_X \to \Psi_J$ denote the TSLS projection onto Ψ_J:

$$Q_J f = \arg\min_{h \in \Psi_J} \|\Pi_{K(J)} T(f - h)\|_{L^2_W}.$$

Assumption 3 *(i)* $\sup_{h \in \Psi_J, \|h\|_{L^2_X} = 1} \tau_J \|\Pi_{K(J)} Th - Th\|_{L^2_W} \leq v_J$ *where* $v_J < 1$ *for all* $J \in \mathcal{T}$ *and* $v_J \to 0$ *as* $J \to \infty$*;*
(ii) $\tau_J \|T(h_0 - \Pi_J h_0)\|_{L^2_W} \leq C_T \|h_0 - \Pi_J h_0\|_{L^2_X}$ *for all* $J \in \mathcal{T}$*;*
(iii) $\|Q_J(h_0 - \Pi_J h_0)\|_\infty \leq C_Q \|h_0 - \Pi_J h_0\|_\infty$ *for all* $J \in \mathcal{T}$.

Denote the "population" sieve variance of $\hat{h}_J(x)$ as $\|\sigma_{x,J}\|_{sd}^2 = L_{J,x}\Omega_J L_{J,x}'$ where $L_{J,x} = (\psi^J(x))'[S_J' G_{b,J}^{-1} S_J]^{-1} S_J' G_{b,J}^{-1}$ and $\Omega_J = \mathbb{E}[u^2 b^{K(J)}(W)(b^{K(J)}(W))']$ with $u = Y - h_0(X)$, $G_{b,J} = \mathbb{E}[b^{K(J)}(W)(b^{K(J)}(W))']$, and $S_J = \mathbb{E}[b^{K(J)}(W)(\psi^J(X))']$. Also let $\|\sigma_{x,J}\|^2 = (\psi^J(x))'[S_J' G_{b,J}^{-1} S_J]^{-1}(\psi^J(x))$, which satisfies $\|\sigma_{x,J}\| \asymp \|\sigma_{x,J}\|_{sd}$ uniformly in x by Assumption 2.

Assumption 4 *(i)* $\underline{c}\tau_J^2 J \leq \inf_{x\in\mathcal{X}} \|\sigma_{x,J}\|^2 \leq \sup_{x\in\mathcal{X}} \|\sigma_{x,J}\|^2 \leq \overline{C}\tau_J^2 J$ *for all* $J \in \mathcal{T}$*;* *(ii)* $\limsup_{J\to\infty} \sup_{x\in\mathcal{X}, J_2\in\mathcal{T}: J_2>J} (\|\sigma_{x,J}\|_{sd}/\|\sigma_{x,J_2}\|_{sd}) < \gamma$.

Assumptions 1(i)(ii) and 2 are standard conditions on the support of X and W and the conditional variance of the errors (see, e.g., Chen and Christensen (2018)) that can be relaxed. Assumption 1(iii) is an identification condition that is generically satisfied under endogeneity (see Andrews (2017)) and is trivially satisfied for nonparametric regression because T reduces to the identity in that case. Assumption 3 is also trivially satisfied for nonparametric regression with $C_T, C_Q = 1$. Assumption 3(i) is imposed to ensure that $\hat{s}_J^{-1}$ is a suitable sample analog of τ_J. Assumption 3(ii) is the usual L^2 "stability condition" imposed in the NPIV literature to derive L^2-norm rates. Assumption 3(iii) is a L^∞-norm analogue used to control the bias in sup-norm. Chen and Christensen (2018) provide a thorough discussion of Assumption 4(i) and derive primitive sufficient conditions for it in the context of nonparametric demand estimation. Assumption 4(ii) says that $\|\sigma_{x,J}\|_{sd}^2$ is increasing in $J \in \mathcal{T}$, uniformly in x. We view this as mild because J increases exponentially over $\mathcal{T}$. Indeed, by Assumption 2 and 4(i) and the fact that $J \asymp 2^{Ld}$ for some $L \in \mathbb{N}$, for any $J, J_2 \in \mathcal{T}$ with $J_2 > J$ we have

$$\sup_{x\in\mathcal{X}} \frac{\|\sigma_{x,J}\|_{sd}}{\|\sigma_{x,J_2}\|_{sd}} \asymp \frac{\tau_J\sqrt{J}}{\tau_{J_2}\sqrt{J_2}} \leq \frac{\tau_{2^{Ld}}}{\tau_{2^{(L+1)d}}} 2^{-d/2} \leq 2^{-d/2} < 1\,.$$

4.2 Main Results: Adaptive Estimation in Sup-norm

We now show $\tilde{J}$ leads to minimax rate-adaptive estimators of both the structural function h_0 and its derivatives. Our results encompass nonparametric regression as a special case.

We first define the parameter space for h_0. Let $B^p_{\infty,\infty}(M)$ denote the Hölder ball of smoothness p and radius M (see Appendix D.3 for a formal definition). For given constants $C_T, C_Q, M > 0$ and $\overline{p} > \underline{p} > \frac{d}{2}$ with $r \geq \lfloor \overline{p} \rfloor + 1$, let $\mathcal{H}^p = \mathcal{H}^p(M, C_T, C_Q)$ denote the subset of $B^p_{\infty,\infty}(M)$ that satisfies Assumption 3(ii)(iii) for any distribution of (X, W, u) satisfying Assumptions 1-4, and let $\mathcal{H} = \bigcup_{p\in[\underline{p},\overline{p}]} \mathcal{H}^p$. For each $h_0 \in \mathcal{H}$, we let $\mathbb{P}_{h_0}$ denote the distribution of $(X_i, Y_i, W_i)_{i=1}^\infty$ where each observation is generated by an IID draw from a distribution of (X, W, u) satisfying Assumptions 1-4 with $Y = h_0(X) + u$.

Theorem 4.1 *Let Assumptions 1-4 hold.*

(i) Suppose the model is mildly ill-posed. Then: there is a universal constant $C_{4.1}$ for which

$$\sup_{p\in[\underline{p},\overline{p}]}\sup_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\left(\|\hat{h}_{\tilde{J}}-h_0\|_\infty > C_{4.1}\left(\frac{\log n}{n}\right)^{\frac{p}{2(p+\varsigma)+d}}\right)\to 0.$$

(ii) Suppose the model is severely ill-posed. Then: there is a universal constant $C_{4.1}$ for which

$$\sup_{p\in[\underline{p},\overline{p}]}\sup_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\left(\|\hat{h}_{\tilde{J}}-h_0\|_\infty > C_{4.1}(\log n)^{-p/\varsigma}\right)\to 0.$$

We now show $\tilde{J}$ also leads to adaptive estimation of derivatives of h_0. Intuitively, estimating the derivative of h_0 inflates convergence rate of the (squared) bias and variance terms by the same factor (a power of J). Therefore, a rate-optimal choice of J for estimating h_0 is also rate-optimal for estimating derivatives of h_0.

Corollary 4.1 *Let Assumptions 1-4 hold and let $a\in(\mathbb{N}_0)^d$ with $0<|a|<\underline{p}$.*

(i) Suppose the model is mildly ill-posed. Then: there is a universal constant $C'_{4.1}$ for which

$$\sup_{p\in[\underline{p},\overline{p}]}\sup_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\left(\|\partial^a\hat{h}_{\tilde{J}}-\partial^a h_0\|_\infty > C'_{4.1}\left(\frac{\log n}{n}\right)^{\frac{p-|a|}{2(p+\varsigma)+d}}\right)\to 0.$$

(ii) Suppose the model is severely ill-posed. Then: there is a universal constant $C'_{4.1}$ for which

$$\sup_{p\in[\underline{p},\overline{p}]}\sup_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\left(\|\partial^a\hat{h}_{\tilde{J}}-\partial^a h_0\|_\infty > C'_{4.1}(\log n)^{-(p-|a|)/\varsigma}\right)\to 0.$$

Remark 4.1 The convergence rates in Theorem 4.1 and Corollary 4.1 are the minimax rates for estimating h_0 and $\partial^a h_0$ under sup-norm loss; see Chen and Christensen (2018). Hence, $\hat{h}_{\tilde{J}}$ and $\partial^a\hat{h}_{\tilde{J}}$ converge at the minimax rate in both the mildly and severely ill-posed cases. Case (i) encompasses nonparametric regression as a special case with $\varsigma=0$. To the best of our knowledge, Theorem 4.1 and Corollary 4.1 are the first results on adaptive estimation in sup-norm for NPIV and, more generally, ill-posed inverse problems with unknown operator.

Remark 4.2 Our procedure requires the B-spline order r to satisfy $r\geq\lfloor p\rfloor+1$ for exact minimax rate adaptivity. If the true p is larger so that $r<\lfloor p\rfloor+1$, then our method is still "adaptive" in the sense that it yields consistent estimates of h_0 and its derivatives without requiring prior knowledge of the true smoothness of h_0 or the strength of the instruments. In this case the data-driven estimators $\hat{h}_{\tilde{J}}$ and $\partial^a\hat{h}_{\tilde{J}}$ will converge at the rates presented

in Theorem 4.1 and Corollary 4.1 with $p = r$. Thus, our procedure should be attractive to applied researchers who often use a relatively low choice of r in applications. For instance, Arellano, Blundell, and Bonhomme (2017) use linear splines ($r = 2$). While in principle our method could be extended to let r become large, known results from approximation theory imply that the basis becomes ill-conditioned (i.e., collinear) as r increases (see, e.g., Lyche (1978) and Scherer and Shadrin (1999)). As a consequence, the resulting procedure would be less numerically stable than with smaller r.

4.3 Main Results: UCBs for h_0

It is known since Low (1997) that it is impossible to construct confidence bands that are simultaneously honest and adaptive over Hölder classes of different smoothness. As is standard following Picard and Tribouley (2000), Giné and Nickl (2010), Bull (2012), Chernozhukov et al. (2014), and many others, we establish coverage guarantees over a "generic" subclass $\mathcal{G}$ of $\mathcal{H}$. To describe $\mathcal{G}$, first note by the discussion in Appendix D.3 that there exists a constant $\overline{B} < \infty$ for which $\sup_{h \in \mathcal{H}^p} \|h - \Pi_J h\|_\infty \leq \overline{B} J^{-\frac{p}{d}}$ holds for all $J \in \mathcal{T}$ and all $p \in [\underline{p}, \overline{p}]$. For any small fixed $\underline{B} \in (0, \overline{B})$ and any $\underline{J} \in \mathcal{T}$, we define

$$\mathcal{G}^p = \left\{ h \in \mathcal{H}^p : \underline{B} J^{-\frac{p}{d}} \leq \|h - \Pi_J h\|_\infty \text{ for all } J \in \mathcal{T} \text{ with } J \geq \underline{J} \right\}, \quad \mathcal{G} = \bigcup_{p \in [\underline{p}, \overline{p}]} \mathcal{G}^p .$$

The class $\mathcal{G}$ is sometimes called a class of "self-similar" functions. Giné and Nickl (2010, 2016) present several results establishing the genericity of $\mathcal{G}$ in $\mathcal{H}$. Loosely speaking, their results say $\mathcal{H}^p \setminus (\cup_{\underline{B}>0, \underline{J} \in \mathcal{T}} \mathcal{G}^p)$ is nowhere dense in $\mathcal{H}^p$ under the norm topology of $\mathcal{H}^p$. Thus, the set of functions in $\mathcal{H}^p$ but not in $\mathcal{G}^p$ for some $\underline{B}$ and $\underline{J}$ is topologically meagre.

We say that a UCB $\{C_n(x) : x \in \mathcal{X}\}$ is *honest* over $\mathcal{G}$ with level α if

$$\liminf_{n \to \infty} \inf_{h_0 \in \mathcal{G}} \mathbb{P}_{h_0}\big(h_0(x) \in C_n(x) \quad \forall \ x \in \mathcal{X}\big) \geq 1 - \alpha , \tag{25}$$

and *adaptive* if for every $\epsilon > 0$ there exists a constant D for which

$$\liminf_{n \to \infty} \inf_{p \in [\underline{p}, \overline{p}]} \inf_{h_0 \in \mathcal{G}^p} \mathbb{P}_{h_0}\left(\sup_{x \in \mathcal{X}} |C_n(x)| \leq D r_n(p) \right) \geq 1 - \epsilon ,$$

where $|\cdot|$ is Lebesgue measure and $r_n(p)$ is the minimax sup-norm rate of estimation over $\mathcal{H}^p$. Let $C_n(x, A)$ denote the UCB from (16) replacing $\hat{A}$ with a constant $A > 0$. Our first main result is that $C_n(x, A)$ is honest and adaptive in the mildly ill-posed case:

Theorem 4.2 *Let Assumptions 1-4 hold and suppose the model is mildly ill-posed. Then: there is a constant $A^* > 0$ (independent of α) such that for all $A \geq A^*$,*

$$\textit{(i)} \quad \liminf_{n\to\infty} \inf_{h_0 \in \mathcal{G}} \mathbb{P}_{h_0}\big(h_0(x) \in C_n(x, A) \quad \forall \ x \in \mathcal{X}\big) \geq 1 - \alpha\,;$$

$$\textit{(ii)} \quad \inf_{p\in[\underline{p},\overline{p}]} \inf_{h_0 \in \mathcal{G}^p} \mathbb{P}_{h_0}\left(\sup_{x\in\mathcal{X}} |C_n(x, A)| \leq C_{4.2}(1 + A)\left(\frac{\log n}{n}\right)^{\frac{p}{2(p+\varsigma)+d}}\right) \to 1\,,$$

where $C_{4.2} > 0$ is a universal constant.

Remark 4.3 Theorem 4.2 shows that our UCBs are honest and adaptive in mildly ill-posed models (where $\tau_J \asymp J^{\varsigma/d}$) for all $\varsigma \geq 0$. Importantly, the researcher doesn't need to know the true instrument strength as measured by ς to implement our procedures.

Remark 4.4 As the mildly ill-posed case nests nonparametric regression as a special case with $\varsigma = 0$, Theorem 4.2 shows that our UCBs are honest and adaptive for general nonparametric regression models with non-Gaussian, heteroskedastic errors.

Remark 4.5 The constant A^* in Theorem 4.2 depends implicitly on $\underline{B}$ and becomes large as $\underline{B} \downarrow 0$, coherent with the findings of Armstrong (2021) for Gaussian white noise models. This constant cannot be chosen in a data-dependent way (i.e., one cannot adapt to unknown $\underline{B}$). In practice, A can actually be quite small to guarantee coverage for a fixed DGP—see the simulations in Section 5. The UCBs in Section 2 replace a fixed constant A by $\hat{A} = \log\log \tilde{J}$, which increases no faster than $\log\log n$. These UCBs therefore have coverage guarantees over $\mathcal{G}$ defined for any small $\underline{B} > 0$ and contract within a $\log\log n$ factor of the minimax rate.

Theorem 4.2 establishes that the UCBs for h_0 in Procedure 2 is honest and adaptive in the mildly ill-posed case. We have found that the UCBs in Procedure 2 perform well in terms of coverage across many simulation designs including the severely ill-posed design in Appendix C. Nevertheless, for the severely ill-posed case, we can only establish valid coverage of the UCBs in Procedure 2 using the critical value $\text{cv}^*(x)$ corresponding to $\tilde{J} = \hat{J}_n$ case, i.e.,

$$\text{cv}^*(x) = z^*_{1-\alpha} + \hat{A} \max\{\theta^*_{1-\hat{\alpha}}\,,\ \tilde{J}^{-\underline{p}/d}/\hat{\sigma}_{\tilde{J}}(x)\}\,. \tag{26}$$

The term $\tilde{J}^{-\underline{p}/d}$ bounds the order of the bias term $\|\Pi_{\tilde{J}} h_0 - h_0\|_\infty$, which accounts for the fact that the optimal choice of J in severely ill-posed models is bias-dominating. This band reduces to the Procedure 2 UCB when $\theta^*_{1-\hat{\alpha}} \geq \tilde{J}^{-\underline{p}/d}/\hat{\sigma}_{\tilde{J}}(x)$ for all x.

Remark 4.6 In our empirical application to estimating the intensive margin and its elasticity, the UCB (16) using critical value (26) reduces to the Procedure 2 band provided $\underline{p} \geq 0.7$. The condition $\underline{p} \geq 0.7$ is naturally satisfied as h_0 is assumed to be differentiable in order to estimate the elasticity.

Let $C_n(x, A)$ denote the UCB (16) with the critical value (26), except replacing $\hat{A}$ with a constant $A > 0$.

Theorem 4.3 *Let Assumptions 1-4 hold and suppose the model is severely ill-posed. Then: there is a constant $A^* > 0$ (independent of α) such that for all $A \geq A^*$,*

$$\textit{(i)} \quad \liminf_{n\to\infty} \inf_{h_0 \in \mathcal{G}} \mathbb{P}_{h_0}\big(h_0(x) \in C_n(x, A) \quad \forall \ x \in \mathcal{X}\big) \geq 1 - \alpha\,;$$

$$\textit{(ii)} \quad \inf_{p\in[\underline{p},\overline{p}]} \inf_{h_0 \in \mathcal{G}^p} \mathbb{P}_{h_0}\left(\sup_{x\in\mathcal{X}} |C_n(x, A)| \leq C_{4.3}(1 + A)(\log n)^{-\underline{p}/\varsigma} \right) \to 1\,,$$

where $C_{4.3} > 0$ is a universal constant.

Our recommended choice $\hat{A} = \log\log \tilde{J}$ ensures that the UCBs are asymptotically valid over $\mathcal{G}$ for any $\underline{B} > 0$ and contract within a $\log\log n$ factor of the minimax rate if the true smoothness is $p = \underline{p}$, and within a $\log n$ factor of the minimax rate otherwise.

Remark 4.7 If the true $p > \underline{p}$, then the factor $\tilde{J}^{-\underline{p}/d}$ is conservative and the UCB does not contract at the minimax rate. This raises the question as to whether it is possible to construct UCBs that are adaptive in severely ill-posed settings. As stated in Chapter 8.3 of Giné and Nickl (2016), the existence of rate-adaptive UCBs implicitly requires the estimation of certain aspects of the unknown function, e.g. smoothness, to be feasible. In mildly ill-posed settings, the condition $h_0 \in \mathcal{G}^p$ is sufficient to ensure that $\tilde{J}$ diverges at the oracle rate $J_0 \asymp (n/\log n)^{d/(2(p+\varsigma)+d)}$. As it turns out, $\tilde{J}$ is sufficiently informative about the unknown smoothness p to facilitate the construction of adaptive UCBs. In severely ill-posed models the oracle choice is $J_0 = (a \log n)^{d/\varsigma}$ for $0 < a < (2C)^{-1}$, which is independent of p. Therefore, the adaptivity of $\tilde{J}$ cannot be used to ascertain information about p. We conjecture that any UCB that is centered around an adaptive estimator that aims to mimic the oracle $\hat{h}_{J_0}$ will likely face the same "identifiability" problem of recovering information about p from J_0.

4.4 Main Results: UCBs for Derivatives

We now present an analogous set of results for data-driven UCBs for derivatives of h_0. Here we require an additional regularity condition similar to Assumption 4(i), which is only

needed for the results in this subsection. Let $\|\sigma^a_{x,J}\|^2 = (\partial^a \psi^J(x))'[S_J' G_{b,J}^{-1} S_J]^{-1}(\partial^a \psi^J(x))$.

Assumption 4 (continued) *(iii) There exist constants $\underline{c}, \overline{C} > 0$ for which $\underline{c}\tau_J^2 J^{1+2|a|/d} \leq \inf_{x\in\mathcal{X}} \|\sigma^a_{x,J}\|^2 \leq \sup_{x\in\mathcal{X}} \|\sigma^a_{x,J}\|^2 \leq \overline{C}\tau_J^2 J^{1+2|a|/d}$ for all $J \in \mathcal{T}$.*

We first present results for the mildly ill-posed case. Let $C_n^a(x, A)$ denote the UCB $C_n^a(x)$ from (19) when $\hat{A}$ is replaced by a constant $A > 0$.

Theorem 4.4 *Let Assumptions 1-4 hold, $|a| < \underline{p}$, and suppose the model is mildly ill-posed. Then: there is a constant $A^* > 0$ (independent of α) such that for all $A \geq A^*$,*

$$\textit{(i)} \quad \liminf_{n\to\infty} \inf_{h_0\in\mathcal{G}} \mathbb{P}_{h_0}\big(\partial^a h_0(x) \in C_n^a(x, A) \quad \forall\ x \in \mathcal{X}\big) \geq 1 - \alpha\,;$$

$$\textit{(ii)} \quad \inf_{p\in[\underline{p},\overline{p}]} \inf_{h_0\in\mathcal{G}^p} \mathbb{P}_{h_0}\left(\sup_{x\in\mathcal{X}} |C_n^a(x, A)| \leq C_{4.4}(1 + A)\left(\frac{\log n}{n}\right)^{\frac{p-|a|}{2(p+\varsigma)+d}}\right) \to 1\,,$$

where $C_{4.4} > 0$ is a universal constant.

Remark 4.8 As the mildly ill-posed case nests nonparametric regression as a special case, our UCBs are honest and adaptive for derivatives of h_0 in general nonparametric regression models with non-Gaussian, heteroskedastic errors.

As in the previous subsection, for the severely ill-posed case, we can only establish valid coverage of the UCB (19) using the critical value $\mathrm{cv}^{a*}(x)$ corresponding to $\tilde{J} = \hat{J}_n$, i.e.,

$$\mathrm{cv}^{a*}(x) = z^{a*}_{1-\alpha} + \hat{A}\max\{\theta^*_{1-\hat{\alpha}}\,,\ \tilde{J}^{(|a|-\underline{p})/d}/\hat{\sigma}^a_{\tilde{J}}(x)\}\,. \tag{27}$$

This band reduces to the Procedure 2′ UCB when $\theta^*_{1-\hat{\alpha}} \geq \tilde{J}^{|a|-\underline{p}/d}/\hat{\sigma}^a_{\tilde{J}}(x)$ for all x, which is the case in our empirical application.

Let $C_n^a(x, A)$ denote the band (19) with critical value (27) when $\hat{A}$ is replaced by a constant $A > 0$.

Theorem 4.5 *Let Assumptions 1-4 hold, $|a| < \underline{p}$, and suppose the model is severely ill-posed. Then: there is a constant $A^* > 0$ (independent of α) such that for all $A \geq A^*$,*

$$\textit{(i)} \quad \liminf_{n\to\infty} \inf_{h_0\in\mathcal{G}} \mathbb{P}_{h_0}\big(\partial^a h_0(x) \in C_n^a(x, A) \quad \forall\ x \in \mathcal{X}\big) \geq 1 - \alpha\,;$$

$$\textit{(ii)} \quad \inf_{p\in[\underline{p},\overline{p}]} \inf_{h_0\in\mathcal{G}^p} \mathbb{P}_{h_0}\left(\sup_{x\in\mathcal{X}} |C_n^a(x, A)| \leq C_{4.5}(1 + A)(\log n)^{(|a|-\underline{p})/\varsigma}\right) \to 1\,,$$

where $C_{4.5} > 0$ is a universal constant.

5 Additional Simulations

In this section we present two additional simulation studies. The first is a nonparametric IV design with a non-monotonic, non-Lipschitz structural function. The second is a very wiggly nonparametric regression design, which shows that $\tilde{J}$ can choose a relatively high-dimensional model when needed. Finally, Appendix C presents a third set of simulations in an empirically calibrated Engel curve design which is severely ill-posed.

5.1 Nonparametric IV Design

This design features a non-monotonic, non-Lipschitz structural function. We first draw (U, V) from a bivariate normal distribution with mean zero, unit variance, and correlation 0.75, and draw $Z \sim N(0,1)$ independent of (U, V). We then set $W = \Phi(Z)$ where $\Phi(\cdot)$ denotes the standard normal CDF, $X = \Phi(D(Z+V) + (1-D)V)$ where D is an independent Bernoulli random variable taking the values 0 and 1 each with probability 0.5, and

$$Y = \sin(4X)\log(X) + U\,. \tag{28}$$

The structural function $h_0(x) = \sin(4x)\log(x)$ is plotted in Figure 3. Note that the derivative of h_0 diverges to $-\infty$ as $x \downarrow 0$. Therefore, h_0 is Hölder continuous with exponent p for any $p < 1$, but not Lipschitz continuous.

For each simulated data set we compute our data-driven estimator $\hat{h}_{\tilde{J}}$ and UCBs from (16). We compare these with estimators and UCBs using deterministic choices of sieve dimensions for $J = 4$, 5, 7, and 11 (the first few dimensions over which our procedure searches). We again use a cubic B-spline basis to approximate h_0 and a quartic B-spline for the reduced form.

The first panel of Table 2 presents the average sup-norm loss of $\hat{h}_{\tilde{J}}$ across simulations. These are of similar magnitude to the loss for deterministic-J estimates with $J = 4$ and 5 and are much smaller than the loss with $J = 7$ and 11. Our data-driven UCBs demonstrate valid but slightly conservative coverage for smaller n and coverage close to nominal coverage for $n = 10000$. Bands with $J = 4$ have poor coverage while bands with $J = 5$ have valid coverage for the smaller sample sizes but under-cover for $n = 10000$. It seems $J = 7$ or $J = 11$ is required to have valid coverage for $n = 10000$ in this design. Note that while our bands are slightly conservative for smaller J, they are only about 10% wider than the $J = 5$ bands, and less than half the width of the $J = 7$ bands.

In Figure 3 we plot data-driven estimates and UCBs for h_0 and its derivative over $[0.01, 0.99]$ for a sample of size 2500, alongside deterministic-J estimates and UCBs. In

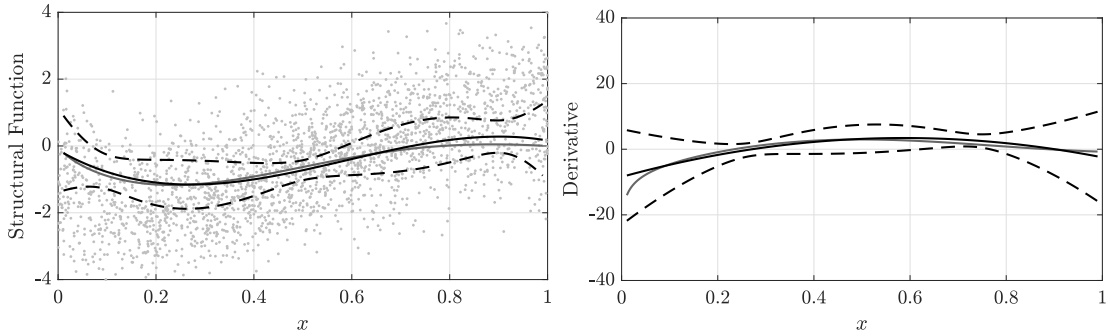
Table 2: Simulation Results for the Nonparametric IV Design (28).

	Data-driven		Deterministic							
			$J=4$		$J=5$		$J=7$		$J=11$	
					Sup-norm Loss					
n	mean	med.	mean	med.	mean	med.	mean	med.	mean	med.
1250	0.541	0.491	0.539	0.489	0.678	0.630	1.087	1.000	1.524	1.422
2500	0.395	0.360	0.393	0.359	0.486	0.451	0.890	0.835	1.342	1.283
5000	0.323	0.292	0.319	0.291	0.367	0.345	0.761	0.696	1.231	1.169
10000	0.262	0.241	0.256	0.239	0.270	0.255	0.623	0.556	1.186	1.136
					UCB Coverage					
	90%	95%	90%	95%	90%	95%	90%	95%	90%	95%
1250	0.997	0.999	0.816	0.892	0.930	0.974	0.951	0.978	0.967	0.984
2500	0.995	0.997	0.744	0.859	0.910	0.950	0.956	0.983	0.978	0.991
5000	0.978	0.992	0.566	0.724	0.881	0.947	0.937	0.976	0.975	0.989
10000	0.908	0.949	0.324	0.470	0.847	0.921	0.935	0.986	0.967	0.989
					95% UCB Relative Width (Deterministic/Data-driven)					
			mean	med.	mean	med.	mean	med.	mean	med.
1250			0.658	0.663	0.925	0.897	1.502	1.451	2.122	2.046
2500			0.661	0.665	0.923	0.908	1.790	1.731	2.554	2.502
5000			0.663	0.668	0.917	0.914	2.255	2.158	3.286	3.228
10000			0.661	0.668	0.913	0.914	2.830	2.757	4.515	4.445

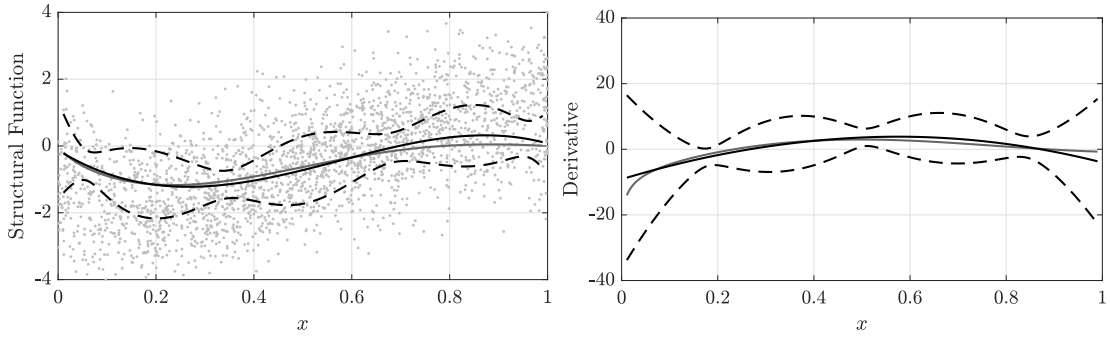
this sample, $\tilde{J} = 4$ and our data-driven UCBs contain the true structural function. The data-driven bands are narrower and more accurately convey the shape of h_0 than the $J = 7$ bands, which are much more wiggly. Our bands are also of a similar width to (but are less wiggly than) the $J = 5$ bands. Panel (d) of Figure 3 also presents data-driven estimates and UCBs for the conditional mean of Y given X. Here the data-driven choice is again $\tilde{J} = 4$. The true structural function falls outside the UCBs for the conditional mean function over almost all of the support of X, highlighting the importance of estimating h_0 using IV methods in this design.

Finally, in Table 3 we present the coverage of our data-driven UCBs $C_n(x, A)$ where we replace $\hat{A} = \log\log \tilde{J}$ with a deterministic choice A ranging over $[0, 1]$. For this design, $A \geq 0.3$ suffices for correct coverage. In particular, $\hat{A} = \log\log \tilde{J}$ yields correct coverage.

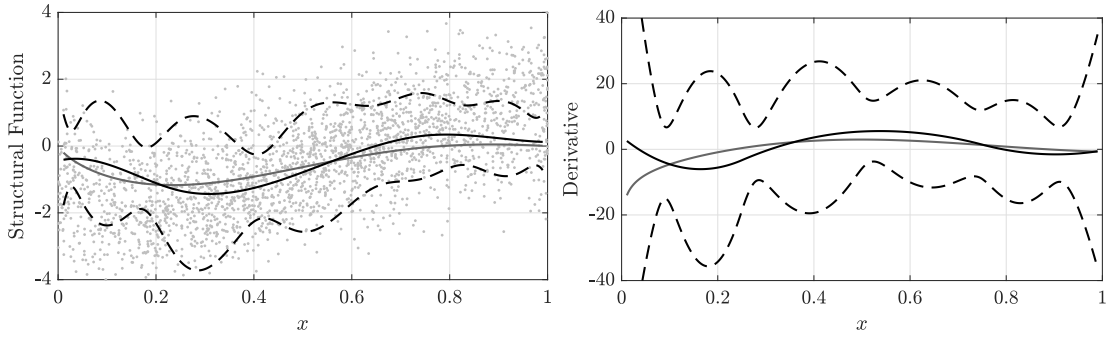
(a) Data-driven Estimates and UCBs



(b) Estimates and UCBs with $J = 5$



(c) Estimates and UCBs with $J = 7$



(d) Data-driven Estimates and UCBs for the Conditional Mean of Y given X

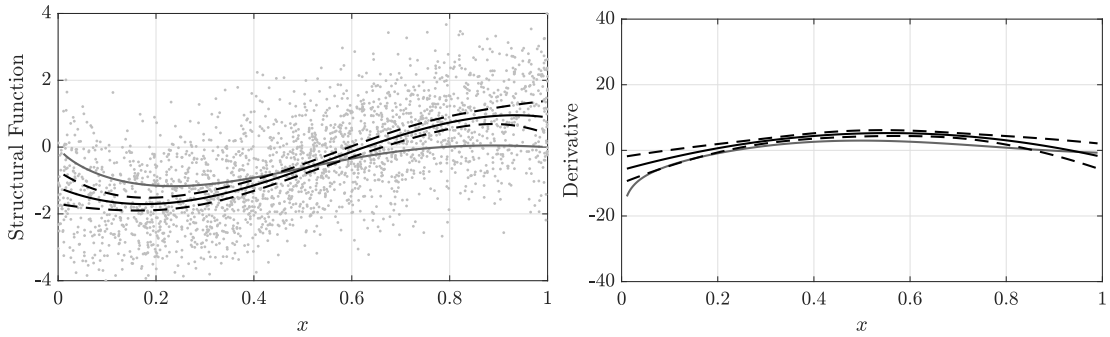


Figure 3: Nonparametric IV design (28): Plots for a sample of size $n = 2500$. Left panels correspond to the structural function, right panels correspond to its derivative. *Note:* Solid grey lines are the true structural function and derivative; solid black lines are estimates, dashed black lines are 95% UCBs. Supports are truncated to $[0.01, 0.99]$ as the derivative is unbounded as $x \downarrow 0$.

Table 3: Coverage of 95% UCBs $C_n(x, A)$, Nonparametric IV Design (28).

	A												
n	0.00	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
1250	0.96	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2500	0.94	0.94	0.95	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5000	0.87	0.88	0.91	0.94	0.97	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00
10000	0.72	0.73	0.78	0.83	0.91	0.95	0.97	0.99	0.99	1.00	1.00	1.00	1.00

5.2 Nonparametric Regression Design

For this design we simulate $X \sim U[0,1]$ and $U \sim N(0,1)$ independently, then set

$$Y = \sin(15\pi X)\cos(X) + U\,. \tag{29}$$

Here $h_0(x) = \sin(15\pi x)\cos(x)$ is very wiggly over $[0,1]$ and requires a high value of J to be selected in order to well approximate h_0 (see Figure 4). While h_0 is infinitely differentiable, its Lipschitz constant is at least 47.1, the Lipschitz constant of its derivative is at least 2220, and Lipschitz constants grow rapidly for higher derivatives.

We again compare our data-driven estimator and UCBs using the procedures described in Appendix A with estimators and UCBs that use deterministic choices of J for $J = 11$, 19, 35, and 67 (these are a subset of values over which our procedure searches). We again use cubic B-splines to approximate h_0.

It is clear from the simulation results presented in Table 4 that $J > 19$ is required to well approximate the true h_0. The average sup-norm loss of $\hat{h}_{\tilde{J}}$ is similar to that of the deterministic-J estimator for $J = 35$, and is smaller than the average loss for all other J presented in the table. Our data-driven UCBs also deliver valid, but conservative, coverage for the true conditional mean function. UCBs based on a deterministic choice of J have zero coverage for $J = 11$ and $J = 19$ as these dimensions are too small to adequately approximate h_0, and tend to under-cover for the remaining J, except perhaps for $J = 35$ when $n = 10000$.

In this design a much smaller value of A suffices to deliver valid coverage, as seen in Table 5. The reason is that the set $\hat{\mathcal{J}}_-$ is large and $\hat{h}_J$ varies a lot across different J due to the wiggliness of h_0. Therefore $z^*_{1-\alpha}$, which is the quantile of a sup-statistic over $\mathcal{X} \times \hat{\mathcal{J}}_-$, is relatively more conservative than for the other designs. This extra conservativeness suffices to deliver valid coverage in this design with smaller A.

Figure 4 plots our data-driven estimator $\hat{h}_{\tilde{J}}$ and 95% UCBs for the conditional mean

Table 4: Simulation Results for the Nonparametric Regression Design (29).

	Data-driven		Deterministic							
			$J = 11$		$J = 19$		$J = 35$		$J = 67$	
					Sup-norm Loss					
n	mean	med.	mean	med.	mean	med.	mean	med.	mean	med.
1250	0.778	0.650	1.242	1.175	0.808	0.732	0.671	0.591	1.111	0.898
2500	0.490	0.423	1.182	1.133	0.705	0.650	0.483	0.415	0.698	0.603
5000	0.347	0.303	1.140	1.109	0.641	0.608	0.332	0.294	0.486	0.426
10000	0.236	0.209	1.113	1.095	0.606	0.585	0.233	0.206	0.330	0.291
					UCB Coverage					
	90%	95%	90%	95%	90%	95%	90%	95%	90%	95%
1250	0.999	0.999	0.000	0.000	0.000	0.000	0.790	0.864	0.627	0.713
2500	1.000	1.000	0.000	0.000	0.000	0.000	0.847	0.899	0.776	0.857
5000	1.000	1.000	0.000	0.000	0.000	0.000	0.857	0.909	0.845	0.910
10000	1.000	1.000	0.000	0.000	0.000	0.000	0.889	0.936	0.867	0.934
			95% UCB Relative Width (Deterministic/Data-driven)							
			mean	med.	mean	med.	mean	med.	mean	med.
1250			0.217	0.209	0.287	0.279	0.410	0.405	0.616	0.582
2500			0.206	0.206	0.279	0.279	0.401	0.405	0.603	0.599
5000			0.190	0.191	0.256	0.260	0.374	0.382	0.565	0.568
10000			0.195	0.196	0.261	0.262	0.380	0.383	0.573	0.572

function for a sample of size 2500. In this sample, $\tilde{J} = 35$. The data-driven estimator well approximates the true conditional mean function h_0, which lies entirely within the 95% UCBs, and the same is true for estimates and UCBs for the derivative of h_0. Deterministic-J bands with $J = 67$ are of a similar width to our data-driven bands for this sample, even though they use a less conservative critical value which only accounts for sampling uncertainty. The estimator is also much wigglier with $J = 67$ than our data-driven estimator and does not approximate h_0 as well.

6 Extensions

So far we have assumed the structural function h_0 is a general d-variate function. As with many other nonparametric estimation problems, minimax rates deteriorate as d increases. This so-called curse of dimensionality applies to any estimator of h_0. However, it can be circumvented by imposing additional structure on h_0 (when appropriate), such as additivity or partial linearity. In this section, we show how our data-driven procedures

Table 5: Coverage of 95% UCBs $C_n(x, A)$, Nonparametric Regression Design (29)

	A												
n	0.00	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
1250	0.88	0.88	0.90	0.93	0.96	0.98	0.98	0.99	0.99	0.99	1.00	1.00	1.00
2500	0.93	0.94	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5000	0.96	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10000	0.97	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

extend to additive and partially linear models.

Additive Structural Functions. Consider first the additive structural function:

$$h_0(x) = c_0 + h_{10}(x_1) + \ldots + h_{d0}(x_d)$$

where $x = (x_1, \ldots, x_d)'$. Here c_0 is a constant representing an "intercept" term and the h_{i0} are suitably normalized for identifiability. In the context of nonparametric regression, Stone (1985) showed that imposing additivity can yield estimators of h_0 that achieve the same (optimal) rate for general d as for $d = 1$.

Our methods may be easily adapted to additive models as follows. We assume for sake of exposition that X is bivariate ($d = 2$). Let $\psi^J(x) = (1, \tilde{\psi}_1^J(x_1)', \tilde{\psi}_2^J(x_2)')'$ where for $i = 1, 2$ we have $\tilde{\psi}_i^J(x_i) = (\tilde{\psi}_{J1}(x_i), \ldots, \tilde{\psi}_{JJ}(x_i))'$. Here J represents the dimensions of sieves used to approximate both h_{10} and h_{20}. The basis functions $\tilde{\psi}_{J1}, \ldots, \tilde{\psi}_{JJ}$ are formed by setting $\tilde{\psi}_{Jj}(x_i) = \psi_{Jj}(x_i) - \int_0^1 \psi_{Jj}(v)\mathrm{d}v$ with $\psi_{J1}(x_1), \ldots, \psi_{JJ}(x_1)$ a univariate B-spline basis. We estimate c_0 and c_J^i, $i = 1, 2$, by TSLS regression of Y on $\psi^J(X)$ using $b^{K(J)}(W)$ as instruments:

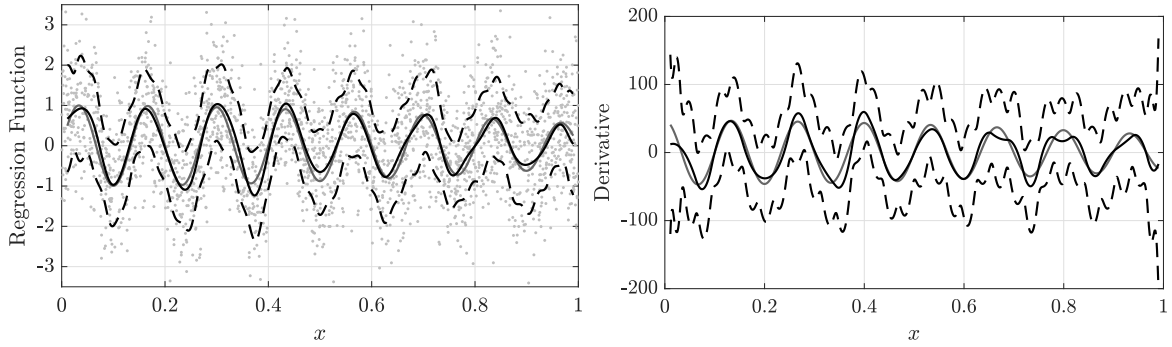
$$\begin{pmatrix} \hat{c}_h \\ \hat{c}_J^1 \\ \hat{c}_J^2 \end{pmatrix} = \left(\mathbf{\Psi}_J' \mathbf{P}_{K(J)} \mathbf{\Psi}_J\right)^- \mathbf{\Psi}_J' \mathbf{P}_{K(J)} \mathbf{Y} = \mathbf{M}_J \mathbf{Y}\,,$$

where the notation is as in Section 2 but with $\psi^J(x) = (1, \tilde{\psi}_1^J(x_1)', \tilde{\psi}_2^J(x_2)')'$. The estimator of h_{i0} is $\hat{h}_{iJ}(x_i) = (\psi_i^J(x_i))'\hat{c}_J^i$. Derivatives of h_{i0} are estimated by differentiating $\hat{h}_{iJ}$.

Our data-driven choice of J is implemented exactly as described in Section 2.3 with $\psi^J(x) = (1, \tilde{\psi}_1^J(x_1)', \tilde{\psi}_2^J(x_2)')'$. Data-driven UCBs for h_{10} are formed analogously to Section 2.4 with two small modifications. First, when computing the critical value $z_{1-\alpha}^*$ in Step 4 of Procedure 2 we now use the sup-statistic

$$\sup_{(x_1, J) \in [0,1] \times \hat{\mathcal{J}}_-} \left| \frac{D_{1J}^*(x_1)}{\hat{\sigma}_{1J}(x_1)} \right|$$

(a) Data-driven Estimates and UCBs



(b) Estimates and UCBs with $J = 67$

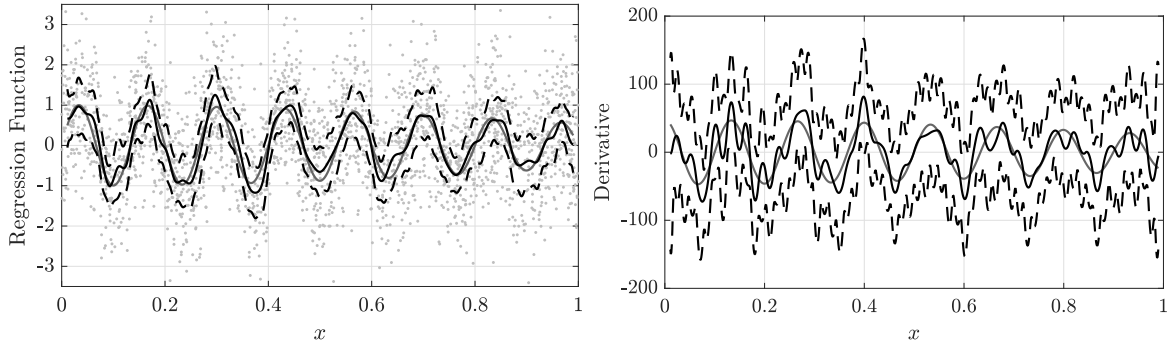


Figure 4: Nonparametric regression design (29): Plots for a sample of size $n = 2500$. Left panels correspond to the conditional mean function, right panels correspond to its derivative. *Note:* Solid grey lines are the true conditional mean function and its derivative; solid black lines are estimates, dashed black lines are 95% UCBs.

where $D^*_{1J}(x_1) = (0, \tilde{\psi}^J_1(x_1)', 0'_J)'\mathbf{M}_J\hat{\mathbf{u}}^*_J$ with 0_J a J-vector of zeros, and

$$\hat{\sigma}^2_{1J}(x) = (0, \tilde{\psi}^J_1(x_1)', 0'_J)\mathbf{M}_J\widehat{\mathbf{U}}_{J,J}\mathbf{M}'_J(0, \tilde{\psi}^J_1(x_1)', 0'_J)'\,.$$

The $100(1-\alpha)\%$ UCB for h_{10} is

$$C_n(x_1) = \left[\hat{h}_{1\tilde{J}}(x_1) - \mathrm{cv}^*(x_1)\hat{\sigma}_{1\tilde{J}}(x_1),\ \hat{h}_{1\tilde{J}}(x_1) + \mathrm{cv}^*(x_1)\hat{\sigma}_{1\tilde{J}}(x_1)\right]$$

with

$$\mathrm{cv}^*(x_1) = \begin{cases} z^*_{1-\alpha} + \hat{A}\theta^*_{1-\hat{\alpha}} & \text{if } \tilde{J} = \hat{J}, \\ z^*_{1-\alpha} + \hat{A}\max\{\theta^*_{1-\hat{\alpha}}\,,\ \tilde{J}^{-\underline{p}}/\hat{\sigma}_{1\tilde{J}}(x_1)\} & \text{if } \tilde{J} = \hat{J}_n \end{cases}$$

where $\underline{p}$ is the minimal smoothness assumed for h_{10} and h_{20}. UCBs for derivatives of h_{10} are constructed analogously.

Partially Linear Structural Functions. An alternative to additivity is the partially linear specification (Ai and Chen, 2003)

$$h_0(x) = h_{10}(x_1) + x_2'\beta_0$$

where x is partitioned as $x = (x_1', x_2')'$ with x_1 of dimension $d_1 < d$, h_{10} is an unknown function, and β_0 is an unknown vector of parameters. When X is exogenous (so $W \equiv X$) this is the important partially linear regression model of Robinson (1988).

Our methods may be adapted to estimate and construct UCBs for h_{10} as follows. First, we let $\psi^J(x) = (\psi_1^J(x_1)', x_2')'$ where $\psi_1^J(x_1) = (\psi_{J1}(x_1), \ldots, \psi_{JJ}(x_1))'$.[14] We estimate c_J and β by TSLS regression of Y on $\psi^J(X)$ using $b^{K(J)}(W)$ as instruments:

$$\begin{pmatrix} \hat{c}_J \\ \hat{\beta} \end{pmatrix} = \left(\mathbf{\Psi}_J' \mathbf{P}_{K(J)} \mathbf{\Psi}_J\right)^- \mathbf{\Psi}_J' \mathbf{P}_{K(J)} \mathbf{Y} = \mathbf{M}_J \mathbf{Y}\,,$$

where the notation is as in Section 2 but with $\psi^J(x) = (\psi_1^J(x_1)', x_2')'$. The estimator of h_{10} is $\hat{h}_{1J}(x_1) = (\psi_1^J(x_1))'\hat{c}_J$. Derivatives of h_{10} are again estimated by differentiating $\hat{h}_{1J}$. When X is exogenous, we simply take $w = x$ and $b^K(w) = \psi^J(x)$.

Our data-driven choice of J is implemented analogously to Section 2.3, except we form the contrasts D_J, $D_J(x) - D_{J_2}(x)$, and $D_J^*(x) - D_{J_2}^*(x)$ and the variance terms $\hat{\sigma}_J^2(x)$ and $\hat{\sigma}_{J,J_2}^2(x)$ using $\psi_0^J(x_1) := (\psi_1^J(x_1)', 0_{d_2})'$ in place of $\psi^J(x)$. As such, the t-statistics are functions of x_1 only and the supremums in the sup-statistics in Steps 2 and 3 of Procedure 1 only need to be computed over the support $\mathcal{X}_1$ of x_1. UCBs for h_{10} are constructed analogously to Section 2.4, where the contrast $D_J^*(x)$ and the variance term $\hat{\sigma}_J^2(x)$ are again formed using $\psi_0^J(x_1)$ in place of $\psi^J(x)$. The $100(1-\alpha)\%$ UCB for h_{10} is

$$C_n(x_1) = \left[\hat{h}_{1\tilde{J}}(x_1) - \mathrm{cv}^*(x_1)\hat{\sigma}_{1\tilde{J}}(x_1),\ \hat{h}_{1\tilde{J}}(x_1) + \mathrm{cv}^*(x_1)\hat{\sigma}_{1\tilde{J}}(x_1)\right]$$

with

$$\mathrm{cv}^*(x_1) = \begin{cases} z_{1-\alpha}^* + \hat{A}\theta_{1-\hat{\alpha}}^* & \text{if } \tilde{J} = \hat{J}, \\ z_{1-\alpha}^* + \hat{A}\max\{\theta_{1-\hat{\alpha}}^*\,,\ \tilde{J}^{-\underline{p}/d_1}/\hat{\sigma}_{1\tilde{J}}(x_1)\} & \text{if } \tilde{J} = \hat{J}_n \end{cases}$$

where $\underline{p}$ is the minimal degree of smoothness assumed for h_{10}. UCBs for derivatives of h_{10} are constructed analogously.

[14] We assume without loss of generality that the X_2 variables have mean zero, which permits identification of h_0 and β. In practice these variables can be de-meaned.

7 Conclusion

We have introduced data-driven procedures for estimation and inference on a nonparametric structural function h_0 and its derivatives using instrumental variables. Our data-driven choice of sieve dimension leads to estimators of h_0 and its derivatives that converge at the fastest possible (i.e., minimax) rate in sup-norm. Our data-driven uniform confidence bands (UCBs) for h_0 and its derivatives are shown to have coverage guarantees and contract at, or within a logarithmic factor of, the minimax rate. Both procedures have good finite sample performance in various simulation designs, including empirically-calibrated trade and Engel curve designs. Our methods are simple to compute, and are applied to estimate and construct UCBs for the elasticity of the intensive margin of firm exports in a monopolistic competition model of international trade.

Aside from the extensions in Section 6, it would be straightforward to extend our methods to weakly dependent data, which is relevant for dynamic causal inference and reinforcement learning. It would also be interesting to consider sup-norm rate-minimaxity jointly with respect to both p and the degree of ill-posedness.

A Nonparametric Regression

Here we specialize our data-driven procedures to nonparametric regression. The conditional mean function $h_0(x) = \mathbb{E}[Y|X = x]$ is estimated by

$$\hat{h}_J(x) = (\psi^J(x))'\hat{c}_J \ , \quad \hat{c}_J = (\mathbf{\Psi}'_J\mathbf{\Psi}_J)^- \mathbf{\Psi}'_J\mathbf{Y} \, .$$

Notation is as in Section 2.3, except now we set $\mathbf{M}_J = (\mathbf{\Psi}'_J\mathbf{\Psi}_J)^-\mathbf{\Psi}'_J$.

1. Compute an upper truncation point $\hat{J}_{\max}$ of the index set as

$$\hat{J}_{\max} = \min\left\{J \in \mathcal{T} : J\sqrt{\log J}\upsilon_n \leq 10\sqrt{n} < J^+\sqrt{\log J^+}\upsilon_n\right\} \tag{30}$$

with $\upsilon_n = \max\{1, (0.1 \log n)^4\}$, then compute $\hat{\mathcal{J}}$ as in (11) with this choice of $\hat{J}_{\max}$.

2. Let $\theta^*_{1-\hat{\alpha}}$ denote the $(1-\hat{\alpha})$ quantile of (12) across independent draws of $(\varpi_i)_{i=1}^n$.

3. Take $\tilde{J} = \hat{J}$ for $\hat{J}$ defined in (13).

Data-driven UCBs are also constructed analogously.

4. For UCBs for h_0, compute the critical value $z^*_{1-\alpha}$ as in (15). For UCBs for $\partial^a h_0$, compute the critical value $z^{a*}_{1-\alpha}$ as in (18).

5. The UCB for h_0 is

$$C_n(x) = \left[\hat{h}_{\tilde{J}}(x) - \left(z^*_{1-\alpha} + \hat{A}\theta^*_{1-\hat{\alpha}}\right)\hat{\sigma}_{\tilde{J}}(x),\ \hat{h}_{\tilde{J}}(x) + \left(z^*_{1-\alpha} + \hat{A}\theta^*_{1-\hat{\alpha}}\right)\hat{\sigma}_{\tilde{J}}(x)\right].$$

The UCB for $\partial^a h_0$ is

$$C^a_n(x) = \left[\partial^a\hat{h}_{\tilde{J}}(x) - \left(z^{a*}_{1-\alpha} + \hat{A}\theta^*_{1-\hat{\alpha}}\right)\hat{\sigma}^a_{\tilde{J}}(x),\ \partial^a\hat{h}_{\tilde{J}}(x) + \left(z^{a*}_{1-\alpha} + \hat{A}\theta^*_{1-\hat{\alpha}}\right)\hat{\sigma}^a_{\tilde{J}}(x)\right].$$

Theorem 4.1 and Corollary 4.1 establish that $\tilde{J}$ leads to estimators of h_0 and its derivatives that attain the minimax sup-norm rates for nonparametric regression. Theoretical properties of the data-driven UCBs are established in Theorems 4.2 and 4.4.

B Additional Details for Section 3

B.1 Basis Functions

We construct basis functions the same way in both the simulations and empirical application. We use cubic B-splines ($r = 4$) to approximate h_0 and quartic B-splines ($r = 5$) to estimate the reduced-form. We also link the dimensions J and $K(J)$ using $q = 2$.

As B-splines are supported on $[0, 1]$ but $\tilde{\pi}_{ij}$ is negative, we transform $\tilde{\pi}_{ij}$ to $[0, 1]$ using $\tilde{\pi} \mapsto \max\{0, \tilde{\pi}/10 + 1\}$. Under this transformation the very small fraction of observations for which $\tilde{\pi}_{ij} < -10$ or, equivalently, $\pi_{ij} < 0.005\%$, are truncated to zero (there were only four such observations in the empirical application). Similarly, we transform z_{ij} to have support $[0, 1]$ using its empirical CDF. The transformed $\tilde{\pi}_{ij}$ is not uniformly distributed on $[0, 1]$ so we place interior knots at its empirical quantiles. The transformed z_{ij} are uniformly distributed on $[0, 1]$ so we place interior knots uniformly between $[0, 1]$.

B.2 Simulations

DGP. Our first simulation design is based on Head et al. (2014). As in Melitz (2003), the only source of firm heterogeneity in their model is productivity. Hence, $r_{ij}(\omega) = e_{ij}(\omega)$, which is assumed to be lognormally distributed. The extensive margin is

$$\log \epsilon(\pi) = \mu + \sigma\sqrt{2}\text{erf}^{-1}(1 - 2\pi), \tag{31}$$

where $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\frac{1}{2}t^2} \, \mathrm{d}t$ is the error function and erf^{-1} is its inverse, and μ and σ^2 are the mean and variance of $\log e_{ij}$. The intensive margin function may be shown to be

$$\log \rho(\pi) = \mu + \frac{\sigma^2}{2} - \log(2\pi) + \log\left(1 + \text{erf}\left(\frac{\sigma^2}{\sqrt{2}} - \text{erf}^{-1}(1 - 2\pi)\right)\right). \tag{32}$$

Its elasticity is

$$\frac{d \log \rho(\pi)}{d \log \pi} = -1 + 2\pi \frac{\exp\left(-\frac{\sigma^2}{\sqrt{2}}\left(\frac{\sigma^2}{\sqrt{2}} - 2\text{erf}^{-1}(1 - 2\pi)\right)\right)}{1 + \text{erf}\left(\frac{\sigma^2}{\sqrt{2}} - \text{erf}^{-1}(1 - 2\pi)\right)}.$$

Our second simulation design is based on the Pareto specification of Chaney (2008). In this design the intensive margin is $\log \rho(\pi) = \rho \log \pi$ and hence its elasticity is constant.

We generate data on z_{ij} by sampling IID with replacement from the empirical distribution of z_{ij}. We then generate data on π_{ij} and $\bar{x}_{ij}$ as follows. For the lognormal design, we estimate two partially linear IV models based on (21) and (22), namely

$$\begin{aligned} \log \epsilon(\pi_{ij}) &= \beta_\epsilon z_{ij} + \delta_i^\epsilon + \zeta_j^\epsilon + e_{ij}^\epsilon, \\ \log \bar{x}_{ij} - \log \rho(\pi_{ij}) &= \beta_\rho z_{ij} + \delta_i^\rho + \zeta_j^\rho + e_{ij}^\rho. \end{aligned}$$

In the first equation, we treat $\log \epsilon(\pi_{ij})$ as the dependent variable using the functional form (31) with $\mu = -2$ and $\sigma = 1.2$. In the second, we treat $\log \bar{x}_{ij} - \log \rho(\pi_{ij})$ as the dependent variable using the functional form (32). We compute the covariance matrix $\hat{\Sigma}$ of the residuals $(\hat{e}_{ij}^\epsilon, \hat{e}_{ij}^\rho)$. We simulate $(e_{ij}^\epsilon, e_{ij}^\rho)$ as independent $N(0, \hat{\Sigma})$ random vectors. Given e_{ij}^ϵ and z_{ij}, we set $\log \epsilon(\pi_{ij}) = 0.875 z_{ij} - 7 + e_{ij}^\epsilon$, then invert $\log \epsilon(\pi_{ij})$ using (31) to obtain $\log \pi_{ij}$. This gives a distribution with support, mean, and variance roughly calibrated to the data used in the application. We then set $\bar{x}_{ij} = \log \rho(\pi_{ij}) - \tilde{\sigma} \kappa^\tau z_{ij} + \delta_i + \zeta_j + e_{ij}^\rho$ using the functional form (32) for $\log \rho$, with $\delta_i = 0$, $\zeta_j = 0$, and with $\tilde{\sigma} = 2.9$ and $\kappa^\tau = 0.36$ as in AAG. We set exporter and importer FEs to zero for $\log \rho$ so that we can compare the effect of first-stage estimation of these FEs on the performance of our procedures.

We generate data for the Pareto design (for which the elasticity of ρ is constant) as described above, except we use $\log \rho(\log \pi) = -0.23 \log \pi$ in place of (32), where the coefficient -0.23 matches AAG's estimate for the constant elasticity specification.[15]

[15] Note that we maintain the same DGP for π as in the lognormal specification. While one could also generate π using the Pareto assumption, this would change the joint distribution of (π_{ij}, z_{ij}), and hence the instrument strength and degree of endogeneity. We keep the distribution fixed across designs so that any difference in results is attributable to the different structural functions $\log \rho$ only.

Simulation Results for the Log-normal Design without Fixed Effects. We first present in Tables 6 and 7 results for estimating $\log \rho$ and the elasticity of ρ in the log-normal design when we treat the FEs δ_i and ζ_j as zero. These results shut down any estimation error that may be introduced by first-stage estimation of the FEs. Overall, the results are very similar to those reported in Table 1 with first-stage estimation of FEs: the sup-norm loss of the data-driven estimators of $\log \rho$ and the elasticity of ρ are similar in magnitude to estimators with deterministic $J = 4$ or $J = 5$, and are several multiples smaller than those with larger J. Coverage of the fixed J UCBs is generally too small when $J = 4, 5$, whereas our data-driven UCBs deliver valid, albeit conservative, coverage. Our data-driven UCBs also demonstrate an improvement in terms of width relative to the deterministic-J UCBs when J is large enough (say $J = 7, 11$) to ensure sufficient coverage. Rejection probabilities of a test of constant elasticity based on our data-driven UCBs for the elasticity of ρ are also similar to those reported in Table 1. Figure 5 presents plots of estimates and UCBs when we treat the FEs as zero using the same sample as Figure 2 (where FEs were estimated in the first-stage). The estimates and UCBs reported in these figures are virtually identical, indicating first-stage estimation of FEs is innocuous.

Simulation Results for the Pareto Design. We now turn to the Pareto design in which $\log \rho$ is linear and hence the elasticity of ρ is constant. For brevity we just present in Table 8 the simulation results for estimating the elasticity of ρ. We adjust for first-stage estimation of exporter and importer FEs, as in the empirical application. The optimal choice is $J = 4$, which is the smallest dimension of a cubic B-spline basis. There is no bias with $J = 4$ because the basis functions span cubic functions. As can be seen, the maximal error in estimating the elasticity of ρ using $\tilde{J}$ is very close to the estimator with fixed $J = 4$. Data-driven UCBs again demonstrate valid but conservative coverage for the elasticity. UCBs with fixed $J = 4$ demonstrate coverage close to (but still slightly under) nominal coverage with $n = 6088$. The UCBs with fixed $J = 4$ are narrower (by about 36%) than our data-driven UCBs as they do not account for potential approximation bias whereas our bands do. Of course, in a real data application the researcher doesn't know whether the true elasticity is constant, and therefore whether the UCBs with fixed $J = 4$ is sufficient to guarantee coverage. Figure 6 presents plots for a representative sample of size 1522, again implementing our procedures as described in Section 3. With our data-driven choice $\tilde{J} = 4$, our nonparametric IV estimate of $\log \rho$ is very close to linear and our estimated elasticity is very close to the true, constant elasticity.

Table 6: Simulation Results for Estimating $\log \rho$, Log-normal Design, no FEs

	Data-driven		Deterministic							
			$J = 4$		$J = 5$		$J = 7$		$J = 11$	
					Sup-norm Loss					
n	mean	med.	mean	med.	mean	med.	mean	med.	mean	med.
761	0.180	0.146	0.166	0.142	0.184	0.159	0.361	0.302	0.670	0.609
1522	0.120	0.099	0.113	0.096	0.125	0.111	0.265	0.218	0.584	0.537
3044	0.088	0.072	0.080	0.069	0.087	0.080	0.196	0.163	0.510	0.468
6088	0.068	0.053	0.058	0.051	0.063	0.058	0.147	0.121	0.456	0.408
					UCB Coverage					
	90%	95%	90%	95%	90%	95%	90%	95%	90%	95%
761	0.992	0.996	0.885	0.937	0.879	0.938	0.909	0.960	0.937	0.972
1522	0.996	0.998	0.898	0.940	0.893	0.944	0.915	0.955	0.964	0.985
3044	0.998	1.000	0.875	0.937	0.903	0.948	0.933	0.964	0.956	0.988
6088	0.999	1.000	0.864	0.936	0.880	0.949	0.913	0.958	0.951	0.984
			95% UCB Relative Width (Deterministic/Data-driven)							
			mean	med.	mean	med.	mean	med.	mean	med.
761			0.660	0.675	0.693	0.695	1.576	1.422	2.606	2.457
1522			0.668	0.681	0.692	0.696	1.784	1.700	3.425	3.168
3044			0.667	0.682	0.683	0.692	1.904	1.855	4.310	4.029
6088			0.663	0.684	0.673	0.691	2.007	2.008	5.637	5.055

References

Adao, R., C. Arkolakis, and S. Ganapati (2020). Aggregate implications of firm heterogeneity: A nonparametric analysis of monopolistic competition trade models. NBER Working Paper no. 28081, dated November 2020.

Adao, R., A. Costinot, and D. Donaldson (2017). Nonparametric counterfactual predictions in neoclassical models of international trade. *American Economic Review 107*(3), 633–689.

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

Andrews, D. W. (2017). Examples of L2-complete and boundedly-complete distributions. *Journal of Econometrics 199*(2), 213–220.

Arellano, M., R. Blundell, and S. Bonhomme (2017). Earnings and consumption dynamics: A nonlinear panel data framework. *Econometrica 85*(3), 693–734.

Armstrong, T. B. (2021). Adaptation bounds for confidence bands under self-similarity. *Bernoulli 27*(2), 1348–1370.

Babii, A. (2020). Honest confidence sets in nonparametric IV regression and other ill-posed models. *Econometric Theory 36*(4), 658–706.

Table 7: Simulation Results for Estimating the Elasticity of ρ, Log-normal Design, no FEs

	Data-driven		Deterministic							
			$J = 4$		$J = 5$		$J = 7$		$J = 11$	
					Sup-norm Loss					
n	mean	med.	mean	med.	mean	med.	mean	med.	mean	med.
761	0.238	0.179	0.204	0.176	0.308	0.272	0.569	0.468	2.000	1.825
1522	0.161	0.128	0.141	0.124	0.213	0.187	0.374	0.334	1.785	1.614
3044	0.131	0.094	0.104	0.091	0.146	0.135	0.279	0.253	1.544	1.362
6088	0.113	0.071	0.075	0.068	0.104	0.096	0.200	0.179	1.361	1.236
					UCB Coverage					
	90%	95%	90%	95%	90%	95%	90%	95%	90%	95%
761	0.993	0.998	0.895	0.939	0.880	0.931	0.894	0.941	0.939	0.974
1522	0.995	0.998	0.878	0.933	0.892	0.933	0.913	0.953	0.956	0.985
3044	0.997	1.000	0.846	0.918	0.891	0.941	0.918	0.958	0.957	0.984
6088	0.993	0.994	0.820	0.894	0.891	0.943	0.915	0.960	0.955	0.984
	Frequency		95% UCB Relative Width (Deterministic/Data-driven)							
	reject		mean	med.	mean	med.	mean	med.	mean	med.
761	0.067		0.633	0.651	0.936	0.932	1.760	1.570	6.464	6.140
1522	0.304		0.640	0.657	0.920	0.918	1.756	1.615	8.407	8.177
3044	0.819		0.642	0.659	0.895	0.903	1.752	1.623	10.406	10.141
6088	0.966		0.636	0.660	0.869	0.895	1.729	1.694	12.832	12.754

Note: Column "reject" reports the proportion of simulations in which constant functions are excluded from our data-driven 95% UCBs for the true elasticity.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics 186*(2), 345–366.

Berry, S. T. and P. A. Haile (2014). Identification in differentiated products markets using market level data. *Econometrica 82*(5), 1749–1797.

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

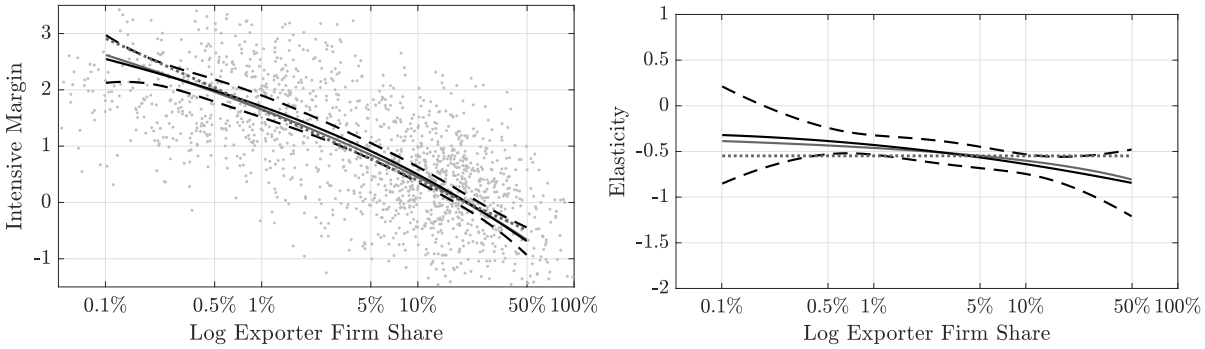
Blundell, R., J. Horowitz, and M. Parey (2017). Nonparametric estimation of a nonseparable demand function under the slutsky inequality restriction. *The Review of Economics and Statistics 99*(2), 291–304.

Breunig, C. and X. Chen (2020). Adaptive, rate-optimal testing in instrumental variables models. arxiv:2006.09587 [econ.em].

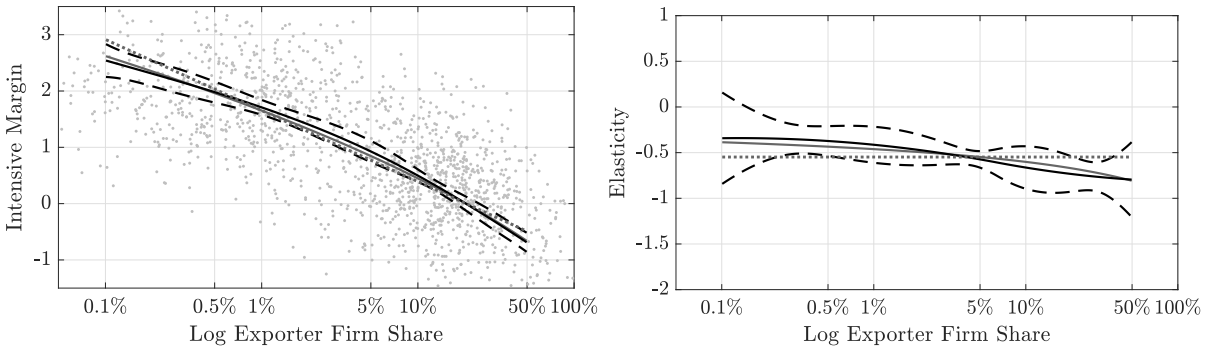
Breunig, C. and X. Chen (2021). Simple adaptive estimation of quadratic functionals in nonparametric iv models. arxiv:2101.12282 [math.st].

Breunig, C. and J. Johannes (2016). Adaptive estimation of functionals in nonparametric instrumental regression. *Econometric Theory 32*(3), 612–654.

(a) Data-driven Estimates and UCBs



(b) Estimates and UCBs with $J = 5$



(c) Estimates and UCBs with $J = 7$

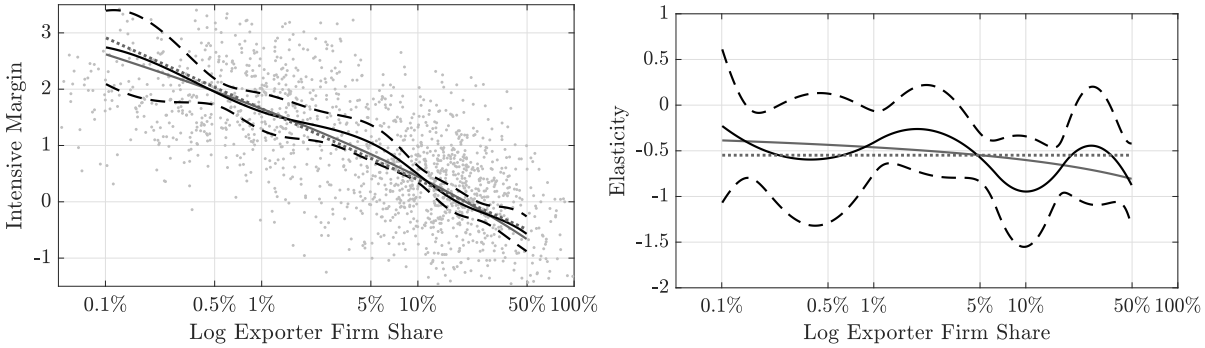


Figure 5: Log-normal design without fixed effects: Plots for a representative sample of size 1522. Left panels correspond to the intensive margin, right panels correspond to its elasticity. *Note:* Solid grey lines are the true curves; solid black lines are estimates; dashed black lines are 95% UCBs; dotted grey lines are linear IV estimates.

Bull, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electronic Journal of Statistics 6*, 1490–1516.

Chaney, T. (2008). Distorted gravity: The intensive and extensive margins of international trade. *American Economic Review 98*(4), 1707–1721.

Chen, X. and T. Christensen (2015a). Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation. arxiv:1508.03365v1 [stat.me].

Chen, X. and T. M. Christensen (2015b). Optimal uniform convergence rates and asymp-

Table 8: Simulation Results for Estimating the Elasticity of ρ, Pareto Design

	Data-driven		Deterministic							
			$J=4$		$J=5$		$J=7$		$J=11$	
			Sup-norm Loss							
n	mean	med.	mean	med.	mean	med.	mean	med.	mean	med.
761	0.228	0.164	0.183	0.157	0.283	0.256	0.519	0.429	1.866	1.699
1522	0.163	0.118	0.125	0.115	0.193	0.172	0.343	0.305	1.620	1.469
3044	0.125	0.085	0.092	0.082	0.134	0.122	0.254	0.229	1.394	1.236
6088	0.095	0.058	0.061	0.056	0.093	0.085	0.180	0.161	1.212	1.092
			UCB Coverage							
	90%	95%	90%	95%	90%	95%	90%	95%	90%	95%
761	0.994	0.998	0.884	0.934	0.854	0.916	0.877	0.926	0.906	0.958
1522	0.993	0.997	0.884	0.937	0.873	0.936	0.892	0.946	0.939	0.976
3044	0.998	1.000	0.878	0.937	0.872	0.932	0.891	0.949	0.940	0.975
6088	0.994	0.998	0.890	0.942	0.886	0.939	0.901	0.954	0.944	0.982
			95% UCB Relative Width (Deterministic/Data-driven)							
			mean	med.	mean	med.	mean	med.	mean	med.
761			0.626	0.650	0.928	0.934	1.760	1.553	6.439	6.149
1522			0.633	0.656	0.909	0.920	1.753	1.619	8.315	8.110
3044			0.637	0.658	0.890	0.906	1.754	1.642	10.373	10.128
6088			0.635	0.660	0.872	0.899	1.738	1.709	12.901	12.862

totic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics 188*(2), 447–465.

Chen, X. and T. M. Christensen (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics 9*(1), 39–84.

Chen, X. and Z. Qi (2022). On well-posedness and minimax optimal rates of nonparametric q-function estimation in off-policy evaluation. *arXiv preprint arXiv:2201.06169*.

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

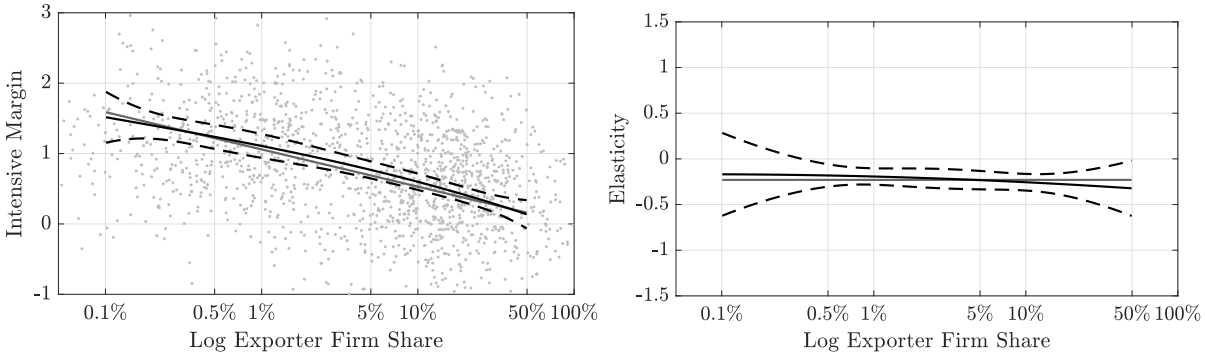
Chen, Y., L. Xu, C. Gulcehre, T. L. Paine, A. Gretton, N. De Freitas, and A. Doucet (2022). On instrumental variable regression for deep offline policy evaluation. *Journal of Machine Learning Research 23*(1).

Chernozhukov, V., D. Chetverikov, and K. Kato (2014). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics 42*(5), 1787–1818.

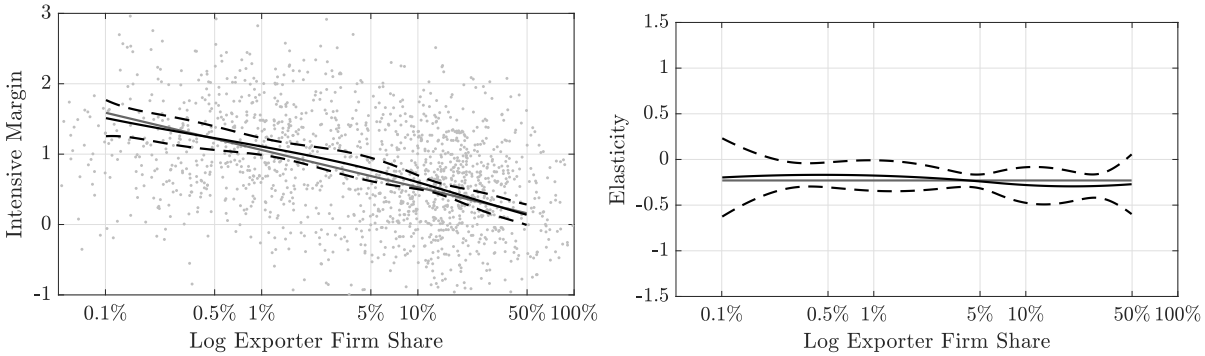
Chernozhukov, V., W. K. Newey, and A. Santos (2023). Constrained conditional moment restriction models. *Econometrica 91*(2), 709–736.

Chetverikov, D. and D. Wilhelm (2017). Nonparametric instrumental variable estimation under monotonicity. *Econometrica 85*(4), 1303–1320.

(a) Data-driven Estimates and UCBs



(b) Estimates and UCBs with $J = 5$



(c) Estimates and UCBs with $J = 7$

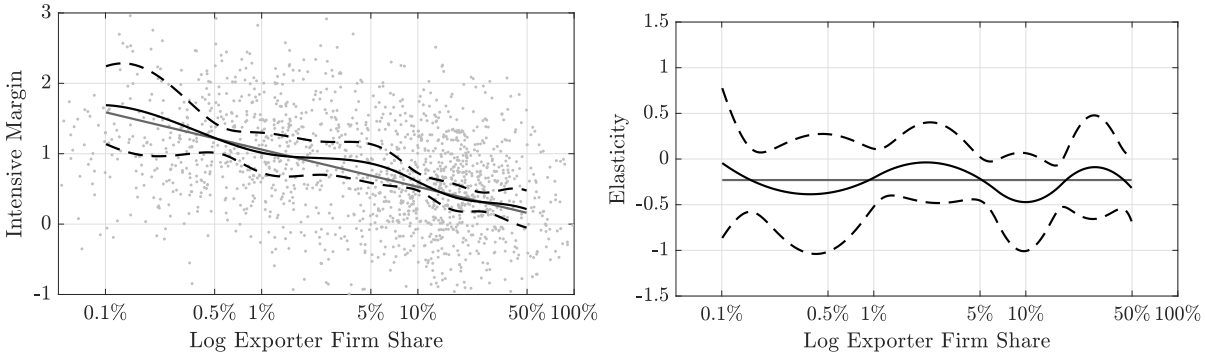


Figure 6: Pareto design (with first-stage estimation of fixed effects): Plots for a representative sample of size 1522. Left panels correspond to the intensive margin, right panels correspond to its elasticity. *Note:* Solid grey lines are the true curves; solid black lines are estimates; dashed black lines are 95% UCBs.

Compiani, G. (2022). Market counterfactuals and the specification of multiproduct demand: A nonparametric approach. *Quantitative Economics 13*(2), 545–591.

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

Eaton, J., S. Kortum, and F. Kramarz (2011). An anatomy of international trade: Evidence from french firms. *Econometrica 79*(5), 1453–1498.

Freyberger, J. and B. Reeves (2019). Inference under shape restrictions. Working paper.

Giné, E. and R. Nickl (2010). Confidence bands in density estimation. *The Annals of Statistics 38*(2), 1122–1170.

Giné, E. and R. Nickl (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press.

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

Head, K., T. Mayer, and M. Thoenig (2014). Welfare and trade without Pareto. *American Economic Review 104*(5), 310–316.

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

Horowitz, J. L. (2014). Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter. *Journal of Econometrics 180*(2), 158–173.

Horowitz, J. L. and S. Lee (2012). Uniform confidence bands for functions estimated nonparametrically with instrumental variables. *Journal of Econometrics 168*(2), 175–188.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics 15*(3), 958–975.

Low, M. G. (1997). On nonparametric confidence intervals. *The Annals of Statistics 25*(6), 2547–2554.

Lyche, T. (1978). A note on the condition numbers of the b-spline bases. *Journal of Approximation Theory 22*(3), 202–205.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica 71*(6), 1695–1725.

Melitz, M. J. and S. J. Redding (2014). Heterogeneous firms and trade. In G. Gopinath, E. Helpman, and K. Rogoff (Eds.), *Handbook of International Economics*, Volume 4, Chapter 1, pp. 1–54. Elsevier.

Melitz, M. J. and S. J. Redding (2015). New trade models, new welfare implications. *American Economic Review 105*(3), 1105–1146.

Miao, W., Z. Geng, and E. J. Tchetgen Tchetgen (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika 105*(4), 987–993.

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

Picard, D. and K. Tribouley (2000). Adaptive confidence interval for pointwise curve estimation. *The Annals of Statistics 28*(1), 298–335.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Scherer, K. and A. Shadrin (1999). New upper bound for the b-spline basis condition number: Ii. a proof of de boor's 2k-conjecture. *Journal of Approximation Theory 99*(2), 217–229.

Spokoiny, V. and N. Willrich (2019). Bootstrap tuning in Gaussian ordered model selection. *The Annals of Statistics 47*(3), 1351–1380.

Stone, C. J. (1985). Additive Regression and Other Nonparametric Models. *The Annals of Statistics 13*(2), 689–705.

Online Appendix to "Adaptive Estimation and Uniform Confidence Bands for Nonparametric Structural Functions and Elasticities"

Xiaohong Chen Timothy Christensen Sid Kankanala

C Additional Simulation: Engel Curves

In this appendix we present additional simulation results for estimating a nonparametric structural function in an empirically calibrated Engel curve setting. The design is based on the British Family Expenditure Survey data used in Blundell et al. (2007). We draw household expenditure X and household income W from a bivariate normal density with correlation $\rho = 0.52$, which is the sample correlation of the expenditure and income data used in Blundell et al. (2007). We then transform X and W to have Uniform$[0,1]$ marginals using their respective inverse marginal CDFs. As a consequence, X and W are linked via a Gaussian copula and the design is severely ill-posed.[16] We then set $h_0(x) = \Phi(5x - 2.5)$ and set $u = h_0(X) - \mathbb{E}[h_0(X)|W] + v$ for $v \sim N(0, 0.01)$. The implementation is the same as the other Monte Carlos from Section 5. For each simulated data set we compute our data-driven estimator $\hat{h}_{\hat{J}}$ and UCBs from (16). We compare these with estimators and UCBs using deterministic choices of sieve dimensions for $J = 4$, 5, 7, and 11 (the first few dimensions over which our procedure searches). We again use a cubic B-spline basis to approximate h_0 and a quartic B-spline basis for the reduced form.

Turning first to the simulation results presented in Table 9, we see that the average sup-norm loss of our data-driven estimator is similar to that of an estimator $\hat{h}_J$ for deterministic J with $J = 4$ and several multiples smaller than that with $J = 5$, 7, or 11. This is to be expected, as the design is severely ill-posed and the true function is very smooth, so a very small choice of J is appropriate. Of course, in practice the researcher does not know the degree of ill-posedness or the degree of smoothness of the structural function.

The second panel of Table 9 shows our data-driven UCBs have valid, albeit conservative, coverage across all sample sizes. By contrast, undersmoothed UCBs with $J = 4$ and $J = 5$ under-cover for $n = 2500$, 5000, and 10000. Undersmoothed UCBs with $J = 7$ have valid but conservative coverage, but these are 40% (with $n = 1250$) to 250% (with $n = 10000$) wider than our data-driven UCBs. It is important to note that although the

[16] This follows from, e.g., Beare (2010), equation (3.3).

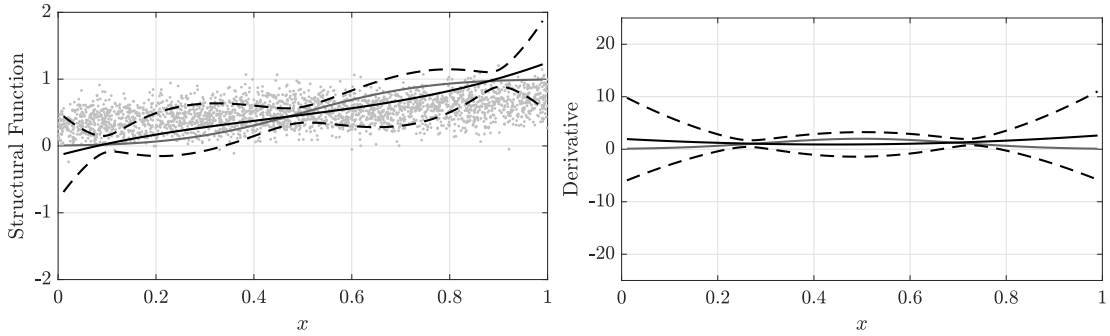
Table 9: Simulation Results for the Engel Curve Design.

	Data-driven		Deterministic							
			$J=4$		$J=5$		$J=7$		$J=11$	
			Sup-norm Loss							
n	mean	med.	mean	med.	mean	med.	mean	med.	mean	med.
1250	0.221	0.183	0.218	0.180	0.337	0.287	0.450	0.400	0.558	0.488
2500	0.167	0.139	0.164	0.138	0.285	0.240	0.417	0.369	0.526	0.468
5000	0.115	0.094	0.113	0.094	0.233	0.197	0.361	0.318	0.484	0.432
10000	0.083	0.068	0.080	0.068	0.173	0.148	0.322	0.299	0.448	0.414
			UCB Coverage							
	90%	95%	90%	95%	90%	95%	90%	95%	90%	95%
1250	0.998	0.999	0.917	0.961	0.903	0.952	0.934	0.972	0.916	0.968
2500	0.998	0.999	0.868	0.931	0.867	0.943	0.950	0.980	0.941	0.982
5000	0.998	0.999	0.833	0.896	0.884	0.939	0.967	0.989	0.968	0.987
10000	0.991	0.994	0.700	0.826	0.826	0.904	0.956	0.988	0.964	0.992
			95% UCB Relative Width (Deterministic/Data-driven)							
			mean	med.	mean	med.	mean	med.	mean	med.
1250			0.653	0.658	1.056	0.978	1.431	1.351	1.798	1.718
2500			0.655	0.661	1.213	1.120	1.797	1.692	2.372	2.270
5000			0.656	0.661	1.458	1.331	2.219	2.107	3.082	2.991
10000			0.658	0.664	1.577	1.478	2.712	2.574	3.962	3.792

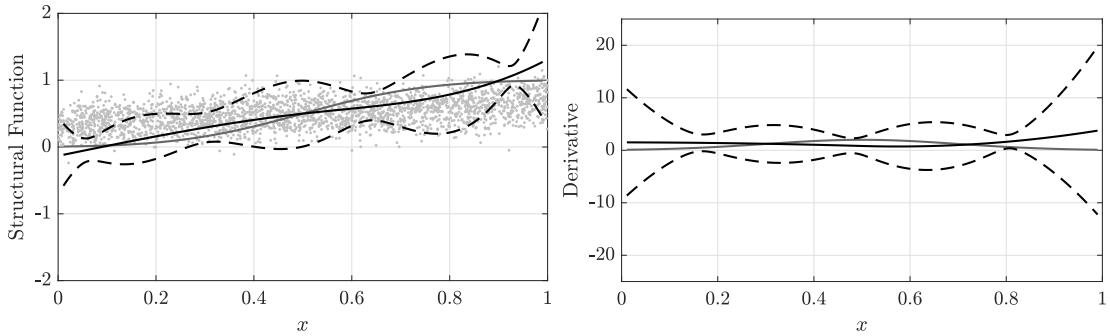
design is severely ill-posed, we are reporting coverage of our UCBs (16). In each simulated data set we have $\hat{J} = \tilde{J}$ irrespective of the sample size n, so the critical value is effectively $z^*_{1-\alpha} + \hat{A}\theta^*_{1-\hat{\alpha}}$. While Theorem 4.2 does not formally establish coverage guarantees of this band in the severely ill-posed case, these simulation results show that the band nevertheless has good coverage in this empirically relevant design.

Figure 7 presents plots of data-driven estimates and UCBs for h_0 and its derivative for a sample of size 2500, alongside deterministic-J estimates and UCBs. In this sample, $\tilde{J} = 4$ and our data-driven UCBs contain the true structural function. As with the other simulations, the data-driven bands are narrower and more accurately convey the shape of h_0 than the $J = 7$ bands, which are much more wiggly. Our bands are also slightly narrower than the $J = 5$ bands. Panel (d) of Figure 3 also presents data-driven estimates and UCBs for the conditional mean of Y given X. Evidently, the true structural function falls outside the UCBs for the conditional mean function over almost all of the support of X, again highlighting the importance of estimating h_0 using IV methods in this design.

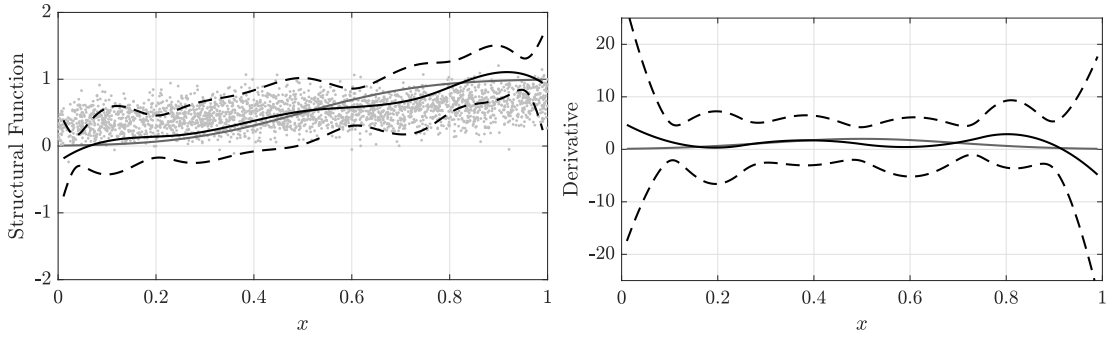
(a) Data-driven Estimates and UCBs



(b) Estimates and UCBs with $J = 5$



(c) Estimates and UCBs with $J = 7$



(d) Data-driven Estimates and UCBs for the Conditional Mean of Y given X

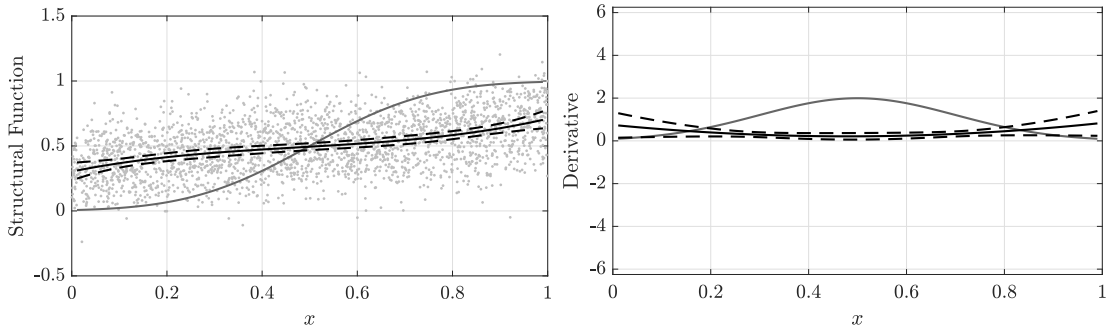


Figure 7: Engel curve design: Plots for a sample of size $n = 2500$. Left panels correspond to the structural function, right panels correspond to its derivative. *Note:* Solid grey lines are the true structural function and derivative; solid black lines are estimates, dashed black lines are 95% UCBs.

D Basis Functions and Hölder Classes

Let Ψ_J denote the closed linear subspace of L_X^2 spanned by a basis $\{\psi_{J1}, \ldots, \psi_{JJ}\}$. We use the following notation for vectors and matrices formed from the basis functions

$$\begin{aligned} \psi_x^J &= (\psi_{J1}(x), \ldots, \psi_{JJ}(x))', & b_w^K &= (b_{K1}(w), \ldots, b_{KK}(w))', \\ \zeta_{\psi,J} &= \sup_{x\in[0,1]^d} \|G_{\psi,J}^{-1/2}\psi_x^J\|_{\ell^2}, & \zeta_{b,J} &= \sup_{w\in[0,1]^{d_w}} \|G_{b,J}^{-1/2}b_w^{K(J)}\|_{\ell^2}, \\ G_{\psi,J} &= \mathbb{E}\big[\psi_X^J(\psi_X^J)'\big], & G_{b,J} &= \mathbb{E}\big[b_W^{K(J)}(b_W^{K(J)})'\big], \\ S_J &= \mathbb{E}\big[b_W^{K(J)}(\psi_X^J)'\big], & S_J^o &= G_{b,J}^{-1/2}\mathbb{E}\big[b_W^{K(J)}(\psi_X^J)'\big]G_{\psi,J}^{-1/2}. \end{aligned}$$

Let s_J be the smallest singular value of $(G_{b,J})^{-1/2}S_J(G_{\psi,J})^{-1/2}$. By Lemma A.1 of Chen and Christensen (2018), under Assumptions 1 and 3(i) there is a finite positive constant a_τ such that

$$a_\tau^{-1}s_J^{-1} \le \tau_J \le s_J^{-1} \quad \text{for all} \quad J \in \mathcal{T}. \tag{33}$$

D.1 B-splines

The construction of univariate B-spline bases supported on $[0,1]$ follows Chapter 12.3 of DeVore and Lorentz (1993). The basis is characterized by an *order* $r \in \mathbb{N}$ (or *degree* $r-1$) and a *resolution level* $l \in \mathbb{N}\cup\{0\}$. Let N_r denote the r-fold convolution of the indicator function of the unit interval, $N_r = \mathbb{1}_{[0,1]} * \cdots * \mathbb{1}_{[0,1]}$ (r-times). A dyadic[17] B-spline basis on $[0,1]$ with resolution level l and order r is

$$\psi_{J_1 j}(x) = N_r(2^l x + r - j), \quad j = 1, \ldots, 2^l + r - 1 =: J_1.$$

In the multivariate case we take tensor products of univariate bases. A B-spline basis supported on $[0,1]^d$ of order r and resolution level l has dimension $J = (2^l + r - 1)^d$. The set of possible sieve dimensions J is therefore $\mathcal{T} = \{(2^l + r - 1)^d : l \in \mathbb{N}\cup\{0\}\}$.

We now review properties of B-spline bases that are used in the technical arguments below. The following Lemma summarizes Lemma E.2 of Chen and Christensen (2018).

Lemma D.1 *Let Assumption 1(i) hold. Then for $\psi^J(x)$ formed from tensor product B-splines, there are constants $C_\psi, a_\zeta > 0$ depending only on a_f such that (i) $\sup_{x\in[0,1]^d}\|\psi^J(x)\|_{\ell^1} \le C_\psi$; (ii) $C_\psi^{-1}J^{-1} \le \lambda_{\min}(G_{\psi,J}) \le \lambda_{\max}(G_{\psi,J}) \le C_\psi J^{-1}$; (iii) $\sqrt{J} \le \zeta_{\psi,J} \le a_\zeta\sqrt{J}$.*

[17] This basis is equivalent to a B-spline basis with interior knots at $2^{-l}, \ldots, 1-2^{-l}$. This knot placement ensures bases are nested across different l (equivalently, J). For irregularly spaced data, interior knots can be placed at the $2^{-l}, \ldots, 1-2^{-l}$ quantiles of the distribution of X.

Corollary D.1 *Let Assumption 1(ii) hold. Then for $b^{K(J)}(w)$ formed from tensor product B-splines and $J \leq K(J) \lesssim J$, there are constants $C_b, a_\zeta > 0$ depending only on a_f such that (i) $\sup_{w\in[0,1]^{d_w}} \|b^{K(J)}(w)\|_{\ell^1} \leq C_b$; (ii) $C_b^{-1} J^{-1} \leq \lambda_{\min}(G_{b,J})) \leq \lambda_{\max}(G_{b,J})) \leq C_b J^{-1}$; (iii) $\sqrt{J} \leq \zeta_{b,J} \leq a_\zeta \sqrt{J}$.*

We also use some continuity properties of B-splines in the proofs. Note that $N_r(\cdot)$ is Lipschitz with $r = 2$ and $r-2$ times continuously differentiable when $r > 2$. Hence, $\|G_{\psi,J}^{-1/2}\big([\psi^J(x_1)]-[\psi^J(x_2)]\big)\|_{\ell^2} \leq CJ^\omega \|x_1 - x_2\|_{\ell^2}^{\omega'}$ holds for some positive constants C, ω, ω'. The B-spline basis also satisfies a Bernstein inequality (or inverse estimate): $\|\partial^a f\|_\infty \lesssim J^{|a|/d}\|f\|_\infty$ holds for any $f \in \Psi_J$ and multi-index a with $|a| < r-1$.

D.2 CDV Wavelets

The construction of CDV wavelet bases supported on $[0,1]$ is reviewed in Appendix E.2 of Chen and Christensen (2018) and follows Cohen, Daubechies, and Vial (1993); see also chapter 4.3.5 of Giné and Nickl (2016). The basis is characterized by an *order* $N \in \mathbb{N}$. Let L denote the smallest integer for which $2^L \geq 2N$. For each *resolution level* $l \geq L$, there are a total of 2^l basis functions. In the multivariate case we generate bases supported on $[0,1]^d$ by taking tensor products of univariate bases. The set of possible J is therefore $\mathcal{T} = \{2^{ld} : l = L+1, L+2, \ldots\}$.

We say that the CDV wavelet basis is S-regular if it is S times continuously differentiable. A S-regular basis can always be chosen by choosing the order N such that $0.18(N-1) \geq S$ (Giné and Nickl, 2016, Theorem 4.2.10(e)). The regularity S of the basis for the endogenous variable X should be chosen such that $S > \overline{p}$, where $\overline{p}$ is the maximal assumed degree of smoothness for h_0. Equivalently, our procedures deliver adaptivity over any smoothness range $[\underline{p}, \overline{p}]$ with $S > \overline{p} > \underline{p} > d/2$ when implemented with a S-regular CDV wavelet basis for X. As with choosing the order r of B-splines, choosing S is analogous to choosing the order of a kernel in kernel-based nonparametric estimation.

CDV wavelet bases for the d_w-dimensional instrumental variable W are constructed similarly, using a basis of regularity $S+1$. Given the resolution level l for the basis for X, the resolution level for the basis for W is $l_w = \lceil (l+q)d/d_w \rceil$ for some $q \in \mathbb{N}$. Linking l_w to l in this manner again defines a mapping $K(J)$ between the two bases that satisfies $\lim_{J\to\infty} K(J)/J = c \in [1, \infty)$. As with B-splines, we recommend that q should be the second- or third-smallest value for which $K(J) \geq J$ holds for all J.

We now review properties of CDV wavelet bases that are used in the proofs below. The following Lemma summarizes Lemma E.4 of Chen and Christensen (2018).

Lemma D.2 *Let Assumption 1(i) hold. Then with $\psi^J(x)$ formed from tensor product CDV wavelets, there are constants $C_\psi, a_\zeta > 0$ depending only on a_f such that (i) $\sup_{x\in[0,1]^d} \|\psi_x^J\|_{\ell^1} \leq C_\psi\sqrt{J}$; (ii) $C_\psi^{-1} \leq \lambda_{\min}(G_{\psi,J}) \leq \lambda_{\max}(G_{\psi,J}) \leq C_\psi$; (iii) $\sqrt{J} \leq \zeta_{\psi,J} \leq a_\zeta\sqrt{J}$.*

Corollary D.2 *Let Assumption 1(ii) hold. Then with $b^{K(J)}(w)$ formed from tensor product CDV wavelets and $J \leq K(J) \lesssim J$, there are constants $C_b, a_\zeta > 0$ depending only on a_f such that (i) $\sup_{w\in[0,1]^{d_w}} \|b_w^{K(J)}\|_{\ell^1} \leq C_b\sqrt{J}$; (ii) $C_b^{-1} \leq \lambda_{\min}(G_{b,J}) \leq \lambda_{\max}(G_{b,J}) \leq C_b$; (iii) $\sqrt{J} \leq \zeta_{b,J} \leq a_\zeta\sqrt{J}$.*

We also use some continuity properties of CDV wavelets in the proofs. As the Daubechies wavelet functions are S times continuously differentiable on their supports, it follows by Lemma D.2(ii) that the basis functions are Hölder continuous, in the sense that $\|G_{\psi,J}^{-1/2}([\psi_{x_1}^J] - [\psi_{x_2}^J])\|_{\ell^2} \leq CJ^\omega\|x_1 - x_2\|_{\ell^2}^{\omega'}$ holds for some positive constants C, ω, ω'. This basis also satisfies a Bernstein inequality (or inverse estimate): $\|\partial^a f\|_\infty \lesssim J^{|a|/d}\|f\|_\infty$ holds for any $f \in \Psi_J$ and multi-index a with $|a| < S$.

D.3 Hölder Classes

Let $B^p_{\infty,\infty} = \{h \in L^\infty([0,1]^d) : \|h\|_{B^p_{\infty,\infty}} < \infty\}$ denote the Hölder space of smoothness p where $\|\cdot\|_{B^p_{\infty,\infty}}$ denotes the Hölder norm of smoothness $p > 0$ (see Giné and Nickl (2016), pp. 370-1), and let $B^p_{\infty,\infty}(M) = \{h \in B^p_{\infty,\infty} : \|h\|_{B^p_{\infty,\infty}} \leq M\}$ denote the Hölder ball of smoothness p and radius M. For $p \notin \mathbb{N}$, we have $h \in B^p_{\infty,\infty}$ if and only if

$$\|h\|_{C^{\lfloor p\rfloor}} + \sum_{a:|a|=\lfloor p\rfloor} \sup_{\substack{x,y\in[0,1]^d:\\ x\neq y}} \frac{|\partial^a h(x) - \partial^a h(y)|}{|x-y|^{p-\lfloor p\rfloor}} < \infty\,,$$

where

$$\|h\|_{C^{\lfloor p\rfloor}} = \|h\|_\infty + \sum_{|a|=\lfloor p\rfloor} \|\partial^a h\|_\infty\,.$$

The space $B^p_{\infty,\infty}$ may equivalently be defined by the error in approximating a function using a linear B-spline basis (see DeVore and Popov (1988) and DeVore and Lorentz (1993)). To do so, let Ψ_J be a CDV wavelet space of regularity $S > p$ or dyadic B-spline space of degree $r - 1 > p$ at resolution level L_J that generates J. Let $d(h, \Psi_J) = \inf_{g\in\Psi_J} \|h - g\|_\infty$. We then have

$$h \in B^p_{\infty,\infty} \iff \|h\|_\infty + \sup_{J:J\in\mathcal{T}} J^{p/d} d(h, \Psi_J) < \infty\,,$$

and, moreover, $\|h\|_\infty + \sup_{J:J\in\mathcal{T}} J^{p/d} d(h,\Psi_J)$ is equivalent to $\|h\|_{B^p_{\infty,\infty}}$; see, e.g., Theorem 12.3.3. of DeVore and Lorentz (1993) for the scalar case and Theorem 4.8 of DeVore and Popov (1988) for the multivariate case. By Lebesgue's lemma (DeVore and Lorentz, 1993, p. 30), we have

$$d(h,\Psi_J) \le \|h - \Pi_J h\|_\infty \le (1 + \|\Pi_J\|_\infty) d(h,\Psi_J)\,,$$

where $\|\Pi_J\|_\infty := \sup_{h:\|h\|_\infty\le 1} \|\Pi_J h\|_\infty$ is the L^∞ norm of the L^2_X projection onto Ψ_J (sometimes referred to as the Lebesgue constant). Huang (2003) and Chen and Christensen (2015) established that $\|\Pi_J\|_\infty \lesssim 1$ under Assumption 1(i) when Ψ_J is spanned by a (tensor product) B-spline or CDV wavelet basis, respectively. Hence,

$$h \in B^p_{\infty,\infty} \iff \|h\|_\infty + \sup_{J:J\in\mathcal{T}} J^{p/d} \|h - \Pi_J h\|_\infty < \infty\,,$$

and $\|h\|_\infty + \sup_{J:J\in\mathcal{T}} J^{p/d}\|h - \Pi_J h\|_\infty$ is equivalent to $\|\cdot\|_{B^p_{\infty,\infty}}$.

E Technical Results and Proofs of Main Results

In this Appendix we first introduce additional notation. We then present technical results and proofs of the main results from Sections 4.2 and 4.3. We finally present technical results and the proofs of main results for Section 4.4.

E.1 Notation

By the discussion in Appendix D, there are finite positive constants a_ζ and a_b such that

$$a_\zeta \ge \zeta_{\psi,J}/\sqrt{J} \ge 1\,, \qquad a_\zeta \ge \zeta_{b,J}/\sqrt{K(J)} \ge 1\,, \qquad a_b \ge K(J)/J\,.$$

For any sequence $(Z_i)_{i=1}^n$ of random vectors and any function g, let $\mathbb{E}_n[g(Z)] = \frac{1}{n}\sum_{i=1}^n g(Z_i)$. Estimators of the matrices defined at the beginning of Appendix D and their orthogonalized versions are

$$\begin{aligned}
\widehat{G}_{\psi,J} &= \mathbb{E}_n\big[\psi^J_X(\psi^J_X)'\big]\,, & \widehat{G}_{b,J} &= \mathbb{E}_n\big[b^{K(J)}_W(b^{K(J)}_W)'\big]\,,\\
\widehat{G}^o_{\psi,J} &= G^{-1/2}_{\psi,J}\mathbb{E}_n\big[\psi^J_X(\psi^J_X)'\big]G^{-1/2}_{\psi,J}\,, & \widehat{G}^o_{b,J} &= G^{-1/2}_{b,J}\mathbb{E}_n\big[b^{K(J)}_W(b^{K(J)}_W)'\big]G^{-1/2}_{b,J}\,,\\
\widehat{S}_J &= \mathbb{E}_n\big[b^{K(J)}_W(\psi^J_X)'\big]\,, & \widehat{S}^o_J &= G^{-1/2}_{b,J}\mathbb{E}_n\big[b^{K(J)}_W(\psi^J_X)'\big]G^{-1/2}_{\psi,J}\,.
\end{aligned}$$

Sieve variances and related terms are

$$\|\hat{\sigma}_{x,J,J_2}\|_{sd}^2 \equiv n\hat{\sigma}_{J,J_2}^2(x) = \|\hat{\sigma}_{x,J}\|_{sd}^2 + \|\hat{\sigma}_{x,J_2}\|_{sd}^2 - 2\hat{\sigma}_{x,J,J_2}\,, \quad \|\hat{\sigma}_{x,J}\|_{sd}^2 \equiv n\hat{\sigma}_J^2(x) = \hat{\sigma}_{x,J,J}\,,$$
$$\|\sigma_{x,J,J_2}\|_{sd}^2 = \|\sigma_{x,J}\|_{sd}^2 + \|\sigma_{x,J_2}\|_{sd}^2 - 2\sigma_{x,J,J_2}\,, \qquad \|\sigma_{x,J}\|_{sd}^2 = \sigma_{x,J,J}\,,$$

where

$$\hat{\sigma}_{x,J,J_2} \equiv n\tilde{\sigma}_{J,J_2}(x) = \hat{L}_{J,x}\widehat{\Omega}_{J,J_2}(\hat{L}_{J_2,x})'\,, \qquad \hat{L}_{J,x} = [\psi_x^J]'[\widehat{S}_J'\widehat{G}_{b,J}^{-1}\widehat{S}_J]^{-1}\widehat{S}_J'\widehat{G}_{b,J}^{-1}\,,$$
$$\sigma_{x,J,J_2} = L_{J,x}\Omega_{J,J_2}(L_{J_2,x})'\,, \qquad L_{J,x} = [\psi_x^J]'[S_J'G_{b,J}^{-1}S_J]^{-1}S_J'G_{b,J}^{-1}\,,$$

with $\hat{\sigma}_{J,J_2}^2(x)$ and $\tilde{\sigma}_{J,J_2}(x)$ given in (9), and

$$\widehat{\Omega}_{J,J_2} = \mathbb{E}_n\left[\hat{u}_J\hat{u}_{J_2}b_W^{K(J)}b_W^{K(J_2)}\right]'\,, \quad \hat{u}_{i,J} = Y_i - \hat{h}_J(X_i)\,, \qquad \widehat{\Omega}_J = \widehat{\Omega}_{J,J}\,,$$
$$\Omega_{J,J_2} = \mathbb{E}\left[u^2b_W^{K(J)}b_W^{K(J_2)}\right]'\,, \quad u_i = Y_i - h_0(X_i)\,, \qquad \Omega_J = \Omega_{J,J}\,.$$

Recall that Π_J is the L_X^2 projection onto Ψ_J. We also define

$$\Delta_J h_0 = h_0 - \Pi_J h_0\,, \qquad \tilde{h}_J(x) = \hat{L}_{J,x}\mathbb{E}_n[b_W^{K(J)}h_0(X)]\,.$$

For bootstrap and related processes, we use the notation

$$\mathbb{Z}_n^*(x,J,J_2) = \frac{1}{\|\hat{\sigma}_{x,J,J_2}\|_{sd}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\hat{L}_{J,x}b_{W_i}^{K(J)}\hat{u}_{i,J} - \hat{L}_{J_2,x}b_{W_i}^{K(J_2)}\hat{u}_{i,J_2}\right)\varpi_i\right)\,, \tag{34}$$

where $(\varpi_i)_{i=1}^n$ are IID $N(0,1)$ draws independent of the data, and

$$\mathbb{Z}_n^*(x,J) \equiv \frac{D_J^*(x)}{\hat{\sigma}_J(x)} = \frac{1}{\|\hat{\sigma}_{x,J}\|_{sd}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n\hat{L}_{J,x}b_{W_i}^{K(J)}\hat{u}_{i,J}\varpi_i\right)\,, \tag{35}$$

$$\widehat{\mathbb{Z}}_n(x,J) = \frac{1}{\|\sigma_{x,J}\|_{sd}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n L_{J,x}b_{W_i}^{K(J)}u_i\varpi_i\right)\,, \tag{36}$$

$$\mathbb{Z}_n(x,J) = \frac{1}{\|\sigma_{x,J}\|_{sd}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n L_{J,x}b_{W_i}^{K(J)}u_i\right)\,. \tag{37}$$

The law of the processes $\mathbb{Z}_n^*(x,J)$ and $\widehat{\mathbb{Z}}_n(x,J)$ is determined from $(\varpi_i)_{i=1}^n$ conditional on the data $\mathcal{Z}^n := (X_i,Y_i,W_i)_{i=1}^n$. We let $\mathbb{P}^*$ denote their probability measure (i.e., with respect to the $(\varpi_i)_{i=1}^n$ conditional on the data) and $\mathbb{E}^*$ denote expectation under $\mathbb{P}^*$. We

also shorten "with $\mathbb{P}_{h_0}$ probability approaching 1 (uniformly over $h_0 \in \mathcal{H}$)" to "wpa1 $\mathcal{H}$-uniformly". We write $\mathcal{H}^p = \mathcal{H} \cap B^p_{\infty,\infty}(M)$ and $\mathcal{G}^p = \mathcal{G} \cap B^p_{\infty,\infty}(M)$.

E.2 Technical Results

Here we present several technical results that are used in the proofs of the main results in Section 4. The proofs of these technical results are presented in our earlier working paper version (Chen, Christensen, and Kankanala, 2022). The following Lemmas E.1 to E.7 are labelled as Lemmas D.1 to D.7 in Chen et al. (2022), whereas the following Theorems E.1 and E.2 are labelled as Theorems D.1 and D.2 in Chen et al. (2022).

We first state two preliminary lemmas used in the proof of Theorem 4.1. The first relates to resolution levels in the mildly ill-posed case. For any positive constant R, define

$$\bar{J}_{\max}(R) = \sup\left\{ J \in \mathcal{T} : J\sqrt{\log J}\left[(\log n)^4 \vee \tau_J\right] \leq R\sqrt{n} \right\}. \tag{38}$$

For $D > 0$ and $p \in [\underline{p}, \overline{p}]$, define

$$\begin{aligned} J_0(p, D) &= \sup\left\{ J \in \mathcal{T} : \tau_J \frac{\sqrt{J}\theta^*_{1-\hat{\alpha}}}{\sqrt{n}} \leq D J^{-\frac{p}{d}} \right\}, \\ J_0^+(p, D) &= \inf\{ J \in \mathcal{T} : J > J_0(p, D) \}. \end{aligned} \tag{39}$$

Lemma E.1 *Let Assumptions 1-4 hold and let $\tau_J \asymp J^{\varsigma/d}$ with $\varsigma \geq 0$. Then: with $\bar{J}_{\max}(R)$ as defined in (38) for any $R > 0$ and $J_0^+(p, D)$ as defined in (39) for any $D > 0$, we have*

$$\inf_{p \in [\underline{p}, \overline{p}]} \inf_{h_0 \in \mathcal{H}^p} \mathbb{P}_{h_0}(J_0^+(p, D) < \bar{J}_{\max}(R)) \to 1.$$

The second lemma relates to resolution levels in the severely ill-posed case. For $R > 0$ and $p \in [\underline{p}, \overline{p}]$, define

$$\bar{J}^*_{\max}(R) = \sup\left\{ J \in \mathcal{T} : \tau_J J\sqrt{\log J} \leq R\sqrt{n} \right\}, \tag{40}$$

$$M_0(p, R) = \sup\{ J \in \mathcal{T} : \tau_J J^{\frac{p}{d}+\frac{1}{2}}\sqrt{\log J} \leq R\sqrt{n} \}, \tag{41}$$

$$M_0^+(p, R) = \inf\{ J \in \mathcal{T} : J > M_0(p, R) \}.$$

Note that $M_0(p, R)$ is (weakly) decreasing in p. In particular, as $\overline{p}/d + 1/2 \geq \underline{p}/d + 1/2 > 1$, we have $\bar{J}^*_{\max}(R) \geq M_0(\underline{p}, R) \geq M_0(p, R) \geq M_0(\overline{p}, R)$ for each R and each $p \in [\underline{p}, \overline{p}]$.

Lemma E.2 *Let $\tau_J \asymp \exp(CJ^{\varsigma/d})$ for some $C, \varsigma > 0$. Then for any $R > 0$, the inequality $M_0^+(\overline{p}, R) \geq J^*_{\max}(R)$ holds for all n sufficiently large.*

E.2.1 Uniform-in-J Convergence Rates for $\hat{h}_J$

Recall the definition of $\bar{J}_{\max}(R)$ from (38) and that $\Delta_J h_0 = h_0 - \Pi_J h_0$.

Theorem E.1 *Let Assumptions 1, 2(i), and 3 hold, and for any positive constant R let $\bar{J}_{\max} \equiv \bar{J}_{\max}(R)$. Then: there exists a universal constant $C_{E.1} > 0$ such that*

$$(i) \quad \inf_{h_0 \in \mathcal{H}} \mathbb{P}_{h_0}\left(\|\tilde{h}_J - h_0\|_\infty \leq C_{E.1} \|\Delta_J h_0\|_\infty \quad \forall \ J \in \mathcal{T} \cap [1, \bar{J}_{\max}] \right) \to 1\,,$$

$$(ii) \quad \inf_{h_0 \in \mathcal{H}} \mathbb{P}_{h_0}\left(\|\hat{h}_J - \tilde{h}_J\|_\infty \leq C_{E.1} \tau_J \frac{\sqrt{J \log \bar{J}_{\max}}}{\sqrt{n}} \quad \forall \ J \in \mathcal{T} \cap [1, \bar{J}_{\max}] \right) \to 1\,.$$

E.2.2 Uniform-in-J Estimation of Sieve Variance Terms

Recall the definition of $\bar{J}_{\max}(R)$ from (38). In the remainder of this subsection, for any fixed $R > 0$, let $\bar{J}_{\max} \equiv \bar{J}_{\max}(R)$. Also let $J_{\min} \to \infty$ arbitrarily slowly. Given $\bar{J}_{\max}$ and $J_{\min}$, define $\mathcal{J}_n = \{J \in \mathcal{T} : J_{\min} \leq J \leq \bar{J}_{\max}\}$,

$$\mathcal{S}_n = \{(x, J, J_2) \in \mathcal{X} \times \mathcal{J}_n \times \mathcal{J}_n : J_2 > J\} \tag{42}$$

and

$$\delta_n = \tau_{\bar{J}_{\max}} \sqrt{\frac{\bar{J}_{\max} \log \bar{J}_{\max}}{n}} + \left(\frac{\bar{J}^2_{\max} \log \bar{J}_{\max}}{n} \right)^{1/3} + J_{\min}^{-\underline{p}/d}\,. \tag{43}$$

Lemma E.3 *Let Assumptions 1-4 hold. Then: there exists universal constants $C_{E.3} > 0$ and $N_{E.3} \in \mathbb{N}$ such that:*

(i) for every $x \in \mathcal{X}$ and $J, J_2 \in \mathcal{T}$ with $J_2 > J \geq N_{E.3}$, we have

$$C_{E.3}^{-1} \|\sigma_{x,J_2}\|_{sd} \leq \|\sigma_{x,J,J_2}\|_{sd} \leq C_{E.3} \|\sigma_{x,J_2}\|_{sd}\,;$$

(ii) we have

$$\inf_{h_0 \in \mathcal{H}} \mathbb{P}_{h_0}\left(\sup_{(x,J,J_2) \in \mathcal{S}_n} \left| \frac{\|\hat{\sigma}_{x,J,J_2}\|_{sd}}{\|\sigma_{x,J,J_2}\|_{sd}} - 1 \right| \leq C_{E.3} \delta_n \right) \to 1\,.$$

Lemma E.4 *Let Assumptions 1-3 hold. Then: there is a universal constant $C_{E.4} > 0$ such that*

$$\inf_{h_0 \in \mathcal{H}} \mathbb{P}_{h_0}\left(\sup_{(x,J,J_2) \in \mathcal{S}_n} \frac{|\hat{\sigma}_{x,J,J_2} - \sigma_{x,J,J_2}|}{\|\sigma_{x,J}\|_{sd} \|\sigma_{x,J_2}\|_{sd}} \leq C_{E.4} \delta_n \right) \to 1\,.$$

In particular,

$$\inf_{h_0\in\mathcal{H}} \mathbb{P}_{h_0}\left(\sup_{(x,J)\in\mathcal{X}\times\mathcal{J}_n} \left| \frac{\|\hat{\sigma}_{x,J}\|_{sd}^2}{\|\sigma_{x,J}\|_{sd}^2} - 1 \right| \leq C_{E.4}\delta_n \right) \to 1\,.$$

E.2.3 Uniform Consistency of $\hat{J}_{\max}$

For the following lemma, recall $\hat{J}_{\max}$ from (10) and $\bar{J}_{\max}(R)$ from (38).

Lemma E.5 *Let Assumptions 1-3 hold. Then: replacing $10\sqrt{n}$ with $M\sqrt{n}$ for any $M > 0$ in the definition of $\hat{J}_{\max}$ from (10), there exists $R_1, R_2 > 0$ which satisfy*

$$\inf_{h_0\in\mathcal{H}} \mathbb{P}_{h_0}\left(\bar{J}_{\max}(R_1) \leq \hat{J}_{\max} \leq \bar{J}_{\max}(R_2) \right) \to 1\,.$$

Remark E.1 *For any $R_2 \geq R_1 > 0$ there exists a finite positive constant C for which*

$$\bar{J}_{\max}(R_1) \leq \bar{J}_{\max}(R_2) \leq C\bar{J}_{\max}(R_1)\,.$$

Lemma E.5 therefore provides an asymptotic rate of divergence for $\hat{J}_{\max}$.

E.2.4 Uniform-in-J Bounds for the Bootstrap

For the following Lemma, recall the critical value $\theta^*_{1-\hat{\alpha}}$ from Section 2.3.

Lemma E.6 *Let Assumptions 1-4 hold. Then: with $\bar{J}_{\max}(R)$ as defined in (38) for any $R > 0$, there exists constants $C_4, C_5 > 0$ for which*

$$\inf_{h_0\in\mathcal{H}} \mathbb{P}_{h_0}\left(C_4\sqrt{\log \bar{J}_{\max}(R)} \leq \theta^*_{1-\hat{\alpha}} \leq C_5\sqrt{\log \bar{J}_{\max}(R)} \right) \to 1.$$

The second is a companion result concerning the critical value $z^*_{1-\alpha}$ from Section 2.4:

Lemma E.7 *Let Assumptions 1-4 hold. Then: with $\bar{J}_{\max}(R)$ as defined in (38) for any $R > 0$, there exists a constant $C_{E.7} > 0$ for which*

$$\inf_{h_0\in\mathcal{H}} \mathbb{P}_{h_0}\left(z^*_{1-\alpha} \leq C_{E.7}\sqrt{\log \bar{J}_{\max}(R)} \right) \to 1\,.$$

E.2.5 Uniform Consistency for the Bootstrap

Recall $\bar{J}_{\max} \equiv \bar{J}_{\max}(R)$ from (38) and $\mathcal{J}_n$ and $\mathcal{S}_n$ from (42).

Theorem E.2 *Let Assumptions 1-4 hold and let* $J_{\min} \asymp (\log \bar{J}_{\max})^2$. *Then: there exists a sequence* $\gamma_n \downarrow 0$ *for which the following inequalities hold wpa1* $\mathcal{H}$*-uniformly:*

$$(i) \quad \sup_{s\in\mathbb{R}} \left| \mathbb{P}_{h_0}\left(\sup_{(x,J)\in\mathcal{X}\times\mathcal{J}_n} \left| \sqrt{n}\frac{\hat{h}_J(x) - \tilde{h}_J(x)}{\|\hat{\sigma}_{x,J}\|_{sd}} \right| \le s \right) - \mathbb{P}^*\left(\sup_{(x,J)\in\mathcal{X}\times\mathcal{J}_n} |\mathbb{Z}_n^*(x,J)| \le s \right) \right| \le \gamma_n \,,$$

$$(ii) \quad \sup_{s\in\mathbb{R}} \left| \mathbb{P}_{h_0}\left(\sup_{(x,J,J_2)\in\mathcal{S}_n} \left| \sqrt{n}\frac{\hat{h}_J(x) - \hat{h}_{J_2}(x) - (\tilde{h}_J(x) - \tilde{h}_{J_2}(x))}{\|\hat{\sigma}_{x,J,J_2}\|_{sd}} \right| \le s \right) \right.$$
$$\left. - \mathbb{P}^*\left(\sup_{(x,J,J_2)\in\mathcal{S}_n} |\mathbb{Z}_n^*(x,J,J_2)| \le s \right) \right| \le \gamma_n \,.$$

E.3 Proofs of Main Results in Sections 4.2 and 4.3

Proof of Theorem 4.1. We first list some constants that will be used throughout the proof. Fix $R_2 > 0$ in the definition of $\bar{J}_{\max}(R_2)$ from (38) sufficiently large so that by Lemma E.5 we have $\inf_{h_0\in\mathcal{H}} \mathbb{P}_{h_0}(\hat{J}_{\max} \le \bar{J}_{\max}(R_2)) \to 1$. Let $\bar{J}_{\max} \equiv \bar{J}_{\max}(R_2)$ for the remainder of the proof. By Theorem E.1(i), there exists $C_{E.1} > 0$ which satisfies

$$\inf_{h_0\in\mathcal{H}} \mathbb{P}_{h_0}\left(\|\tilde{h}_J - \Pi_J h_0\|_\infty \le C_{E.1}\|\Pi_J h_0 - h_0\|_\infty \quad \forall\, J \in [1, \bar{J}_{\max}] \cap \mathcal{T} \right) \to 1. \tag{44}$$

For our choice of sieves, there exists $B_2 > 0$ which satisfies

$$\sup_{p\in[\underline{p},\bar{p}]} \sup_{h_0\in\mathcal{H}^p} J^{\frac{p}{d}} \|\Pi_J h_0 - h_0\|_\infty \le B_2 \quad \forall\ J \in \mathcal{T} \,. \tag{45}$$

Let $\hat{\mathcal{S}} = \{(x, J, J_2) \in \mathcal{X} \times \hat{\mathcal{J}} \times \hat{\mathcal{J}} : J_2 > J\}$. Lemmas E.3 and E.5, Assumption 4(i), and the fact that $\delta_n \downarrow 0$ (cf. (43)) imply that there exists $C_2, C_3 > 0$ which satisfy

$$\inf_{h_0\in\mathcal{H}} \mathbb{P}_{h_0}\left(\sup_{(x,J,J_2)\in\hat{\mathcal{S}}} \frac{\tau_{J_2}\sqrt{J_2}}{\|\hat{\sigma}_{x,J,J_2}\|_{sd}} \le C_3 \right) \to 1 \,, \quad \inf_{h_0\in\mathcal{H}} \mathbb{P}_{h_0}\left(\sup_{(x,J,J_2)\in\hat{\mathcal{S}}} \frac{\|\hat{\sigma}_{x,J,J_2}\|_{sd}}{\tau_{J_2}\sqrt{J_2}} \le C_2 \right) \to 1 \,. \tag{46}$$

Additionally, by Lemma E.6 there exists constants $C_4, C_5 > 0$ which satisfy

$$\inf_{h_0\in\mathcal{H}} \mathbb{P}_{h_0}\left(C_4\sqrt{\log \bar{J}_{\max}} \le \theta^*_{1-\hat{\alpha}} \le C_5\sqrt{\log \bar{J}_{\max}} \right) \to 1. \tag{47}$$

Part (i), step 1: We verify that $\hat{J}$ achieves the optimal rate under mild ill-posedness. Note by the procedure in Appendix A this is sufficient for adaptivity of $\tilde{J}$ for nonparametric regression. Fix $\xi > 1$ ($\xi = 1.1$ in the main text). Choose $D > 0$ such that

$2B_2(C_1+1)D^{-1}C_3 < (\xi-1)$. Recall $J_0(p,D)$ and $J_0^+(p,D)$ from (39); we drop dependence of these quantities on (p,D) hereafter to simplify notation. By Lemma E.1, $\inf_{p\in[\underline{p},\bar{p}]}\inf_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}(J_0^+ < \bar{J}_{\max}) \to 1$. It then follows from Lemmas E.1 and E.5 that $\inf_{p\in[\underline{p},\bar{p}]}\inf_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}(J_0^+ < \hat{J}_{\max}) \to 1$. We therefore assume for the remainder of the proof of part (i) that $J_0^+ < \hat{J}_{\max}, \bar{J}_{\max}$.

By Lemma E.5, $\hat{\mathcal{J}} \subseteq \mathcal{J}_n := \{J \in \mathcal{T} : 0.1(\log \bar{J}_{\max})^2 \le J \le \bar{J}_{\max}\}$ wpa1 $\mathcal{H}$-uniformly. Then for all $J \in \hat{\mathcal{J}}$ with $J > J_0^+$, by the triangle inequality, displays (44) and (45), and definition of J_0, we may deduce that

$$\begin{aligned}
&\left|\|\hat{h}_J - \hat{h}_{J_0^+}\|_\infty - \|\hat{h}_J - \hat{h}_{J_0^+} - (\tilde{h}_J - \tilde{h}_{J_0^+})\|_\infty\right| \\
&\le \|\tilde{h}_J - \Pi_J h_0\|_\infty + \|\tilde{h}_{J_0^+} - \Pi_{J_0^+} h_0\|_\infty + \|\Pi_{J_0^+} h_0 - h_0\|_\infty + \|\Pi_J h_0 - h_0\|_\infty \\
&\le 2B_2(1+C_1)(J_0^+)^{-p/d} \\
&\le 2B_2(1+C_1)D^{-1}\theta^*_{1-\hat{\alpha}}\tau_{J_0^+}\sqrt{J_0^+/n}
\end{aligned}$$

wpa1 uniformly for $h_0 \in \mathcal{H}^p$ and $p \in [\underline{p},\bar{p}]$. By (46), we have that for all $J \in \hat{\mathcal{J}}$ with $J > J_0^+$

$$\tau_{J_0^+}\sqrt{J_0^+} \le \tau_J\sqrt{J} \le C_3\|\hat{\sigma}_{x,J_0^+,J}\|_{sd} \qquad \forall \quad x \in \mathcal{X}$$

wpa1 uniformly for $h_0 \in \mathcal{H}^p$ and $p \in [\underline{p},\bar{p}]$. Combining the preceding two inequalities and using the definition of D, we obtain that for all $J \in \hat{\mathcal{J}}$ with $J > J_0^+$,

$$\sup_{x\in\mathcal{X}}\sqrt{n}\frac{|\hat{h}_J(x) - \hat{h}_{J_0^+}(x)|}{\|\hat{\sigma}_{x,J_0^+,J}\|_{sd}} \le \sup_{x\in\mathcal{X}}\sqrt{n}\frac{|\hat{h}_J(x) - \hat{h}_{J_0^+}(x) - (\tilde{h}_J(x) - \tilde{h}_{J_0^+}(x))|}{\|\hat{\sigma}_{x,J_0^+,J}\|_{sd}} + (\xi-1)\theta^*_{1-\hat{\alpha}}$$

wpa1 uniformly for $h_0 \in \mathcal{H}^p$ and $p \in [\underline{p},\bar{p}]$. It follows by definition of $\hat{J}$ that

$$\begin{aligned}
&\sup_{p\in[\underline{p},\bar{p}]}\sup_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\big(\hat{J} > J_0^+\big) \\
&\le \sup_{p\in[\underline{p},\bar{p}]}\sup_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\left(\sup_{J\in\hat{\mathcal{J}}:J>J_0^+}\sup_{x\in\mathcal{X}}\frac{\sqrt{n}|\hat{h}_{J_0^+}(x) - \hat{h}_J(x)|}{\|\hat{\sigma}_{x,J_0^+,J}\|_{sd}} > \xi\theta^*_{1-\hat{\alpha}}\right) \\
&\le \sup_{h_0\in\mathcal{H}}\mathbb{P}_{h_0}\left(\sup_{(x,J,J_2)\in\hat{\mathcal{S}}}\frac{\sqrt{n}|\hat{h}_J(x) - \hat{h}_{J_2}(x) - (\tilde{h}_J(x) - \tilde{h}_{J_2}(x))|}{\|\hat{\sigma}_{x,J,J_2}\|_{sd}} > \theta^*_{1-\hat{\alpha}})\right) + o(1)\,. \qquad (48)
\end{aligned}$$

To control the r.h.s. probability in (48), let $\hat{\mathcal{J}}(\tilde{J}) = \{J \in \mathcal{T} : 0.1(\log \tilde{J})^2 \le J \le \tilde{J}\}$, $\hat{\mathcal{S}}(\tilde{J}) = \{(x,J,J_2) \in \mathcal{X}\times\hat{\mathcal{J}}(\tilde{J})\times\hat{\mathcal{J}}(\tilde{J}) : J_2 > J\}$, and $\theta^*_{1-\hat{\alpha};\tilde{J}}$ denote the $(1 - 0.5 \wedge$

$\sqrt{(\log \tilde{J})/\tilde{J}})$ quantile of $\sup_{(x,J,J_2)\in\hat{\mathcal{S}}(\tilde{J})} |\mathbb{Z}_n^*(x,J,J_2)|$. Then by Lemma E.5 and Theorem E.2(ii), we have

$$\begin{aligned}
&\sup_{h_0\in\mathcal{H}} \mathbb{P}_{h_0}\bigg(\sup_{(x,J,J_2)\in\hat{\mathcal{S}}} \frac{\sqrt{n}|\hat{h}_J(x) - \hat{h}_{J_2}(x) - (\tilde{h}_J(x) - \tilde{h}_{J_2}(x))|}{\|\hat{\sigma}_{x,J,J_2}\|_{sd}} > \theta^*_{1-\hat{\alpha}} \bigg) \\
&\leq \sup_{h_0\in\mathcal{H}} \sum_{\tilde{J}\in\mathcal{T}:\tilde{J}=\bar{J}_{\max}(R_1)}^{\bar{J}_{\max}(R_2)} \mathbb{P}_{h_0}\bigg(\sup_{(x,J,J_2)\in\hat{\mathcal{S}}(\tilde{J})} \frac{\sqrt{n}|\hat{h}_J(x) - \hat{h}_{J_2}(x) - (\tilde{h}_J(x) - \tilde{h}_{J_2}(x))|}{\|\hat{\sigma}_{x,J,J_2}\|_{sd}} > \theta^*_{1-\hat{\alpha};\tilde{J}} \bigg) \\
&\leq \sum_{\tilde{J}\in\mathcal{T}:\tilde{J}=\bar{J}_{\max}(R_1)}^{\bar{J}_{\max}(R_2)} \Big(\sqrt{(\log \tilde{J})/\tilde{J}} + \gamma_n + o(1) \Big) \to 0\,, \qquad (49)
\end{aligned}$$

where the final line holds for all n large, because $\bar{J}_{\max}(R_1) \to \infty$, $\gamma_n \downarrow 0$, and, by our choice of sieve and Remark E.1, for some constant $C > 0$ we have

$$\begin{aligned}
\#\{J \in \mathcal{T} : \bar{J}_{\max}(R_1) \leq J \leq \bar{J}_{\max}(R_2)\} &\leq \#\{J \in \mathcal{T} : \bar{J}_{\max}(R_1) \leq J \leq C\bar{J}_{\max}(R_1)\} \\
&\leq \#\{l \in \mathbb{N} : \bar{J}_{\max}(R_1) \leq 2^{ld} \leq C\bar{J}_{\max}(R_1)\} \leq C\,.
\end{aligned}$$

In view of (48), this proves $\hat{J} \leq J_0^+$ wpa1 uniformly for $h_0 \in \mathcal{H}^p$ and $p \in [\underline{p}, \overline{p}]$.

Whenever $\hat{J} \leq J_0^+ < \hat{J}_{\max}, \bar{J}_{\max}$, it follows by definition of $\hat{J}$ and display (46) that wpa1 uniformly for $h_0 \in \mathcal{H}^p$ and $p \in [\underline{p}, \overline{p}]$, we have

$$\begin{aligned}
\|\hat{h}_{\hat{J}} - h_0\|_\infty &\leq \|\hat{h}_{\hat{J}} - \hat{h}_{J_0^+}\|_\infty + \|\hat{h}_{J_0^+} - h_0\|_\infty \\
&\leq C_2\xi\theta^*_{1-\hat{\alpha}}\tau_{J_0^+}\sqrt{J_0^+/n} + \|\hat{h}_{J_0^+} - \tilde{h}_{J_0^+}\|_\infty + \|\tilde{h}_{J_0^+} - h_0\|_\infty.
\end{aligned}$$

Then by Theorem E.1, definition of J_0^+, and the lower bound on $\theta^*_{1-\hat{\alpha}}$ in display (47), we may deduce that there exists a constant $C > 0$ for which

$$\inf_{p\in[\underline{p},\overline{p}]} \inf_{h_0\in\mathcal{H}^p} \mathbb{P}_{h_0}\Big(\|\hat{h}_{\hat{J}} - h_0\|_\infty \leq C\theta^*_{1-\hat{\alpha}}\tau_{J_0^+}\sqrt{J_0^+/n} \Big) \to 1.$$

As the model is mildly ill-posed, there exists a constant $C' > 0$ for which $\tau_{J_0^+}\sqrt{J_0^+} \leq C'\tau_{J_0}\sqrt{J_0}$. It then follows by definition of J_0 that

$$\inf_{p\in[\underline{p},\overline{p}]} \inf_{h_0\in\mathcal{H}^p} \mathbb{P}_{h_0}\left(\|\hat{h}_{\hat{J}} - h_0\|_\infty \leq CC'DJ_0^{-p/d} \right) \to 1. \qquad (50)$$

By the upper bound on $\theta^*_{1-\hat{\alpha}}$ in display (47) and because $\sqrt{\log \bar{J}_{\max}} \asymp \sqrt{\log n}$ (as the

model is mildly ill-posed), there exists a constant $E > 0$ such that by defining

$$J_n^*(p, E) = \sup\left\{J \in \mathcal{T} : \tau_J \sqrt{(J \log n)/n} \leq E J^{-p/d}\right\}$$

we have $\inf_{p\in[\underline{p},\overline{p}]} \inf_{h_0\in\mathcal{H}^p} \left(J_n^*(p, E) \leq J_0(p, D)\right) \to 1$. Hence, as $\tau_J \asymp J^{\varsigma/d}$ we have $J_n^*(p, E) \asymp (n/\log n)^{d/(2(p+\varsigma)+d)}$. The desired result now follows from (50).

Part (i), step 2: We verify that $\tilde{J}$ achieves the optimal rate under mild ill-posedness. By step 1, we have $\inf_{p\in[\underline{p},\overline{p}]} \inf_{h_0\in\mathcal{H}^p} \mathbb{P}_{h_0}(\hat{J} \leq J_0^+) \to 1$. If we can show that $\hat{J}_n > J_0^+$ wpa1 $\mathcal{H}$-uniformly, then $\tilde{J} = \hat{J}$ wpa1 $\mathcal{H}$-uniformly and the result follows by step 1.

By the lower bound on $\theta^*_{1-\hat{\alpha}}$ in display (47) and the fact that $\sqrt{\log \bar{J}_{\max}} \asymp \sqrt{\log n}$ (as the model is mildly ill-posed), we may deduce that there exists a constant $E' > 0$ such that $\inf_{p\in[\underline{p},\overline{p}]} \inf_{h_0\in\mathcal{H}^p} \left(J_n^\dagger(p, E') \geq J_0^+(p, D)\right) \to 1$ where

$$J_n^\dagger(p, E') = \inf\left\{J \in \mathcal{T} : \tau_J \sqrt{(J \log n)/n} > E' J^{-p/d}\right\}.$$

But note that $\max_{p\in[\underline{p},\overline{p}]} J_0^\dagger(p, E') = J_0^\dagger(\underline{p}, E')$. The result now follows by Lemma E.5, noting that $\bar{J}_{\max}(R_1)/J_0^\dagger(\underline{p}, E') \to \infty$ when the model is mildly ill-posed because $\underline{p} > d/2$.

Part (ii), step 1: We verify that $\hat{J}_n$ achieves the optimal rate under severe ill-posedness. To simplify notation we assume a CDV wavelet basis, though a similar argument applies (albeit with more complicated notation) for B-splines. Note that when the model is severely ill-posed, for any $R > 0$ we have $n^\beta \lesssim \tau_{\bar{J}_{\max}(R)}$ for some $\beta > 0$ and so $\tau_{\bar{J}_{\max}(R)} > (\log n)^4$ for all sufficiently large n. Therefore $\bar{J}_{\max}(R) = \bar{J}^*_{\max}(R)$ for all n sufficiently large, where $\bar{J}^*_{\max}(R)$ is defined in (40). By Theorem E.1, Lemma E.5, and Remark E.1, we may deduce that there exist constants $D, D' > 0$ for which

$$\begin{aligned}
\|\hat{h}_{\hat{J}_n} - h_0\|_\infty &\leq \|\hat{h}_{\hat{J}_n} - \tilde{h}_{\hat{J}_n}\|_\infty + \|\tilde{h}_{\hat{J}_n} - h_0\|_\infty \\
&\leq D\left((2^{-d}\bar{J}^*_{\max}(R_1))^{-\frac{p}{d}} + \tau_{2^{-d}\bar{J}^*_{\max}(R_2)}\sqrt{2^{-d}\bar{J}^*_{\max}(R_2)\log(2^{-d}\bar{J}^*_{\max}(R_2))/n}\right) \\
&\leq D'\left((2^{-d}\bar{J}^*_{\max}(R_2))^{-\frac{p}{d}} + \tau_{2^{-d}\bar{J}^*_{\max}(R_2)}\sqrt{2^{-d}\bar{J}^*_{\max}(R_2)\log(2^{-d}\bar{J}^*_{\max}(R_2))/n}\right)
\end{aligned}$$

wpa1 uniformly over $\mathcal{H}^p$ and $p \in [\underline{p}, \overline{p}]$.

Recall the definition of $M_0(p, R_2)$ from (41). By Lemma E.2, for all $p \in [\underline{p}, \overline{p}]$ we have that $M_0(p, R_2) \geq M_0(\overline{p}, R_2) \geq 2^{-d} J^*_{\max}(R_2)$ holds for all n sufficiently large, in which case

by definition of $M_0(p, R_2)$ we must have

$$\tau_{2^{-d}\bar{J}^*_{\max}(R_2)}\sqrt{2^{-d}\bar{J}^*_{\max}(R_2)\log(2^{-d}\bar{J}^*_{\max}(R_2))/n} \le R_2(2^{-d}\bar{J}^*_{\max}(R_2))^{-\frac{p}{d}}\,.$$

Combining the preceding two inequalities then yields

$$\|\hat{h}_{\hat{J}_n} - h_0\|_\infty \le D'(1+R_2)2^p(\bar{J}^*_{\max}(R_2))^{-\frac{p}{d}}$$

wpa1 uniformly over $\mathcal{H}^p$ and $p \in [\underline{p}, \overline{p}]$.

It remains to show $(\log n)^{d/\varsigma} \lesssim \bar{J}^*_{\max}(R_2)$ when $\tau_J \asymp \exp(CJ^{\varsigma/d})$ for $C, \varsigma > 0$. Suppose $\liminf_{n\to\infty} \bar{J}^*_{\max}(R_2)/(\log n)^{d/\varsigma} = 0$. Then along a subsequence $\{n_k\}_{k\ge1}$ we have $\bar{J}^*_{\max}(R_2) = (2^{-\varsigma}C^{-1}u_{n_k}\log n_k)^{d/\varsigma}$ for some sequence $u_{n_k} \downarrow 0$. Then $2^d\bar{J}^*_{\max}(R_2) \in \mathcal{T}$ satisfies

$$\tau_{2^d\bar{J}^*_{\max}(R_2)}2^d\bar{J}^*_{\max}(R_2)\sqrt{\log(2^d\bar{J}^*_{\max}(R_2))/n_k} \lesssim n_k^{u_{n_k}^{-\frac{1}{2}}}(\log n_k)^{d/\varsigma}\sqrt{\log\log n_k} \xrightarrow[k\to\infty]{} 0\,,$$

thereby contradicting the definition of $\bar{J}^*_{\max}(R_2)$ from (40) for all sufficiently large k.

Part (ii), step 2: We verify that $\tilde{J}$ achieves the optimal rate under severe ill-posedness. For any constant $D > 0$, by definition of $\tilde{J}$ we have

$$\begin{aligned}
&\sup_{p\in[\underline{p},\overline{p}]}\sup_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\big(\|\hat{h}_{\tilde{J}} - h_0\|_\infty > D(\log n)^{-p/\varsigma}\big)\\
&\le \sup_{p\in[\underline{p},\overline{p}]}\sup_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\big(\|\hat{h}_{\hat{J}} - h_0\|_\infty > D(\log n)^{-p/\varsigma} \text{ and } \hat{J} < \hat{J}_n\big)\\
&\quad + \sup_{p\in[\underline{p},\overline{p}]}\sup_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\big(\|\hat{h}_{\hat{J}_n} - h_0\|_\infty > D(\log n)^{-p/\varsigma}\big).
\end{aligned}$$

By part (ii), step 1, the constant D can be chosen sufficiently large so that the second term on the r.h.s. is $o(1)$. For the first term, note that $\|\hat{h}_{\hat{J}} - h_0\|_\infty \le \|\hat{h}_{\hat{J}} - \hat{h}_{\hat{J}_n}\|_\infty + \|\hat{h}_{\hat{J}_n} - h_0\|_\infty$, so it suffices to show that there exists a constant $D > 0$ for which

$$\sup_{p\in[\underline{p},\overline{p}]}\sup_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\big(\|\hat{h}_{\hat{J}} - \hat{h}_{\hat{J}_n}\|_\infty > D(\log n)^{-p/\varsigma} \text{ and } \hat{J} < \hat{J}_n\big) \to 0\,.$$

But by definition of $\hat{J}$ and displays (46) and (47), we have

$$\begin{aligned}
&\sup_{p\in[\underline{p},\overline{p}]}\sup_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\big(\|\hat{h}_{\hat{J}}-\hat{h}_{\hat{J}_n}\|_\infty > D(\log n)^{-p/\varsigma} \text{ and } \hat{J}<\hat{J}_n\big)\\
&\leq \sup_{p\in[\underline{p},\overline{p}]}\sup_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\left(\xi C_2\theta^*_{1-\hat{\alpha}}\tau_{\hat{J}_n}\sqrt{\hat{J}_n/n} > D(\log n)^{-p/\varsigma}\right)+o(1)\\
&\leq \sup_{p\in[\underline{p},\overline{p}]}\mathbb{1}\left[\xi C_2C_5\tau_{2^{-d}\bar{J}^*_{\max}(R_2)}\sqrt{2^{-d}\bar{J}^*_{\max}(R_2)\log(2^{-d}\bar{J}^*_{\max}(R_2))/n} > D(\log n)^{-p/\varsigma}\right]+o(1)\,.
\end{aligned}$$

By step 1, we have $\tau_{2^{-d}\bar{J}^*_{\max}(R_2)}\sqrt{2^{-d}\bar{J}^*_{\max}(R_2)\log(2^{-d}\bar{J}^*_{\max}(R_2))/n}\lesssim(\log n)^{-p/\varsigma}$ uniformly for $p\in[\underline{p},\overline{p}]$, so the constant D can be chosen sufficiently large that the indicator function on the r.h.s. is zero uniformly for $p\in[\underline{p},\overline{p}]$ for all n sufficiently large. ■

Proof of Corollary 4.1. Part (i): Recall $J_0^+\equiv J_0(p,D)^+$ from (39). We have

$$\|\partial^a\hat{h}_{\hat{J}}-\partial^a h_0\|_\infty\leq\|\partial^a\hat{h}_{\hat{J}}-\partial^a\hat{h}_{J_0^+}\|_\infty+\|\partial^a\hat{h}_{J_0^+}-\partial^a\tilde{h}_{J_0^+}\|_\infty+\|\partial^a\tilde{h}_{J_0^+}-\partial^a h_0\|_\infty\,.$$

As $\hat{J}\leq J_0^+<\hat{J}_{\max},\bar{J}_{\max}$ holds wpa1 uniformly for $h_0\in\mathcal{H}^p$ and $p\in[\underline{p},\overline{p}]$, by part (i), step 1 of the proof of Theorem 4.1, we may appeal to a Bernstein inequality (or inverse estimate) for our choice of basis to write

$$\|\partial^a\hat{h}_{\hat{J}}-\partial^a h_0\|_\infty\lesssim(J_0^+)^{|a|/d}\left(\|\hat{h}_{\hat{J}}-\hat{h}_{J_0^+}\|_\infty+\|\hat{h}_{J_0^+}-\tilde{h}_{J_0^+}\|_\infty\right)+\|\partial^a\tilde{h}_{J_0^+}-\partial^a h_0\|_\infty\,.$$

By similar arguments to the proof of Corollary 3.1 of Chen and Christensen (2018), we may also deduce $\|\partial^a\tilde{h}_{J_0^+}-\partial^a h_0\|_\infty\lesssim(J_0^+)^{(|a|-p)/d}$ and so

$$\|\partial^a\hat{h}_{\hat{J}}-\partial^a h_0\|_\infty\lesssim(J_0^+)^{|a|/d}\left(\|\hat{h}_{\hat{J}}-\hat{h}_{J_0^+}\|_\infty+\|\hat{h}_{J_0^+}-\tilde{h}_{J_0^+}\|_\infty+(J_0^+)^{-p/d}\right)\,.$$

It now follows by similar arguments to part (i), step 1 of the proof of Theorem 4.1 and definition of J_0 that there exists a constant $C>0$ for which

$$\inf_{p\in[\underline{p},\overline{p}]}\inf_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\left(\|\partial^a\hat{h}_{\hat{J}}-\partial^a h_0\|_\infty\leq CJ_0^{(|a|-p)/d}\right)\to1.$$

The result follows from noting, as in the proof of part (i), step 1 of the proof of Theorem 4.1, that

$$\inf_{p\in[\underline{p},\overline{p}]}\inf_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}\big(J_n^*(p,E)\leq J_0(p,D)\big)\to1\,,$$

where $J_n^*(p,E)\asymp(n/\log n)^{d/(2(p+\varsigma)+d)}$, and by part (i), step 2 of the proof of Theorem 4.1

(which shows that $\tilde{J} = \hat{J}$ wpa1 $\mathcal{H}$-uniformly).

Part (ii): Recall $\bar{J}^*_{\max}(R)$ from (40) and $\hat{J}_n$ from the definition of $\tilde{J}$. By similar arguments to part (ii), step 1 of the proof of Theorem 4.1, and the proof of Corollary 3.1 of Chen and Christensen (2018), we may deduce

$$\begin{aligned}&\|\partial^a \hat{h}_{\hat{J}_n} - \partial^a h_0\|_\infty \\ &\lesssim (\bar{J}^*_{\max}(R_2))^{\frac{|a|}{d}} \left((2^{-d}\bar{J}^*_{\max}(R_2))^{-\frac{p}{d}} + \tau_{2^{-d}\bar{J}^*_{\max}(R_2)} \sqrt{2^{-d}\bar{J}^*_{\max}(R_2) \log(2^{-d}\bar{J}^*_{\max}(R_2))/n} \right)\end{aligned}$$

wpa1 uniformly over $\mathcal{H}^p$ and $p \in [\underline{p}, \overline{p}]$. Hence, by part (ii), step 1 of the proof of Theorem 4.1,

$$\|\partial^a \hat{h}_{\hat{J}_n} - \partial^a h_0\|_\infty \lesssim (\log n)^{(|a|-p)/d}$$

wpa1 uniformly over $\mathcal{H}^p$ and $p \in [\underline{p}, \overline{p}]$.

By similar arguments to part (ii), step 2 of the proof of Theorem 4.1, it suffices to show that there exists a constant $C > 0$ for which

$$\sup_{p \in [\underline{p}, \overline{p}]} \sup_{h_0 \in \mathcal{H}^p} \mathbb{P}_{h_0}\big(\|\partial^a \hat{h}_{\hat{J}} - \partial^a \hat{h}_{\hat{J}_n}\|_\infty > C(\log n)^{(|a|-p)/\varsigma} \text{ and } \hat{J} < \hat{J}_n \big) \to 0 \,.$$

But for any $\hat{J} \leq \hat{J}_n$ by a Bernstein inequality (or inverse estimate) for our choice of basis,

$$\|\partial^a \hat{h}_{\hat{J}} - \partial^a \hat{h}_{\hat{J}_n}\|_\infty \lesssim (\hat{J}_n)^{|a|/d} \|\hat{h}_{\hat{J}} - \hat{h}_{\hat{J}_n}\|_\infty \lesssim (\bar{J}^*_{\max}(R_2))^{|a|/d} \|\hat{h}_{\hat{J}} - \hat{h}_{\hat{J}_n}\|_\infty$$

wpa1 uniformly over $\mathcal{H}^p$ and $p \in [\underline{p}, \overline{p}]$, where the second inequality is because $\hat{J}_n \leq \hat{J}_{\max} \leq \bar{J}_{\max}(R_2)$ wpa1 $\mathcal{H}$-uniformly by Lemma E.5 and because $\bar{J}_{\max}(R_2) = \bar{J}^*_{\max}(R_2)$ for all n sufficiently large. But note by severe ill-posedness and definition of $\bar{J}^*_{\max}(R_2)$, we have that $C(\bar{J}^*_{\max}(R_2))^{\varsigma/d} \asymp \log \tau_{\bar{J}^*_{\max}(R_2)} \leq \log(R_2\sqrt{n}) \asymp \log n$, and so $\bar{J}^*_{\max}(R_2) \lesssim (\log n)^{d/\varsigma}$. The result now follows by part (ii), step 2 of the proof of Theorem 4.1. ■

Proof of Theorem 4.2. In some of what follows, we use the fact that the sieve dimensions for CDV wavelet bases are linked via $J^+ = 2^d J$ for $J \in \mathcal{T}$. We do so for notational convenience; a similar argument (with more complicated notation) applies for B-splines.

Part (i), step 1: By part (i), step 2 of the proof of Theorem 4.1, we have $\hat{J} = \tilde{J}$ wpa1 $\mathcal{H}$-uniformly. It therefore suffices to prove the result for the band

$$C_n(x) = \left[\hat{h}_{\hat{J}}(x) - \left(z^*_{1-\alpha} + \hat{A}\theta^*_{1-\hat{\alpha}} \right) \hat{\sigma}_{\hat{J}}(x),\ \hat{h}_{\hat{J}}(x) + \left(z^*_{1-\alpha} + \hat{A}\theta^*_{1-\hat{\alpha}} \right) \hat{\sigma}_{\hat{J}}(x) \right],$$

(cf. (16)). Note by Appendix A this implies the result holds for our UCBs for nonparamet-

ric regression as well. Fix $R_2 > 0$ in the definition of $\bar{J}_{\max}(R_2)$ from (38) sufficiently large so that by Lemma E.5 we have $\inf_{h_0 \in \mathcal{H}} \mathbb{P}_{h_0}(\hat{J}_{\max} \leq \bar{J}_{\max}(R_2)) \to 1$. Let $\bar{J}_{\max} \equiv \bar{J}_{\max}(R_2)$ for the remainder of the proof. Recall the constants $C_{E.1}$ from (44), $\underline{B}$ and $\overline{B}$ from the discussion preceding the statement of this theorem, and C_4 and C_5 from (47). Also note that by Lemmas E.3 and E.5, Assumption 4(i), and the fact that $\delta_n \downarrow 0$ (cf. (43)) imply that there exists $C_2, C_3 > 0$ which satisfy

$$\inf_{h_0 \in \mathcal{H}} \mathbb{P}_{h_0}\left(\sup_{(x,J) \in \mathcal{X} \times \hat{\mathcal{J}}} \frac{\tau_J \sqrt{J}}{\|\hat{\sigma}_{x,J}\|_{sd}} \leq C_3 \right) \to 1 \,, \quad \inf_{h_0 \in \mathcal{H}} \mathbb{P}_{h_0}\left(\sup_{(x,J) \in \mathcal{X} \times \hat{\mathcal{J}}} \frac{\|\hat{\sigma}_{x,J}\|_{sd}}{\tau_J \sqrt{J}} \leq C_2 \right) \to 1 \,. \tag{51}$$

Let $v = \inf_{J \in \mathcal{T}} (1 + \|\Pi_J\|_\infty)^{-1} > 0$, where $\|\Pi_J\|_\infty \lesssim 1$ is the Lebesgue constant for Ψ_J (see Appendix D.3). Choose $\beta \in (0,1)$ and $E > 0$ such that $(v\underline{B}\beta^{-\underline{p}/d} - (C_{E.1}+1)\overline{B}) > 0$ and $E^{-1}(v\underline{B}\beta^{-\underline{p}/d} - (C_{E.1}+1)\overline{B}) > C_2(\xi+1)$, where $\xi > 1$ ($\xi = 1.1$ in the main text).

Define $J_0(p,E)$ as in (39). Part (i), step 1 of the proof of Theorem 4.1 implies that $J_0(p,E) \gtrsim (n/\log n)^{d/(2(p+\varsigma)+d)}$. By Lemma E.5 and mild ill-posedness, for any constant $C > 0$ we have $J_0(p,E)/(\log \hat{J}_{\max})^2 \geq C$ wpa1 uniformly for $h_0 \in \mathcal{H}^p$ and $p \in [\underline{p}, \overline{p}]$. Hence, $\inf\{J \in \mathcal{T} : J \geq \beta J_0(p,E)\} > \log \hat{J}_{\max}$ wpa1 uniformly for $h_0 \in \mathcal{H}^p$ and $p \in [\underline{p}, \overline{p}]$.

Fix any $J \in \hat{\mathcal{J}}$ with $J < \beta J_0(p,E)$ (this is justified wpa1 uniformly for $h_0 \in \mathcal{H}^p$ and $p \in [\underline{p}, \overline{p}]$ by the preceding paragraph) and note (dropping dependence of J_0 on (p,E))

$$\begin{aligned}
\|\hat{h}_J - \hat{h}_{J_0}\|_\infty &= \|\hat{h}_J - \hat{h}_{J_0} - \tilde{h}_J + \tilde{h}_J - \tilde{h}_{J_0} + \tilde{h}_{J_0} - h_0 + h_0\|_\infty \\
&\geq \|\tilde{h}_J - h_0\|_\infty - \|\tilde{h}_{J_0} - h_0\|_\infty - \|\hat{h}_J - \tilde{h}_J - (\hat{h}_{J_0} - \tilde{h}_{J_0})\|_\infty .
\end{aligned}$$

For a given $h_0 \in \mathcal{G}^p$, let $h_{0,J} \in \arg\min_{h \in \Psi_J} \|h - h_0\|_\infty$. Recall $\underline{J}$ from the definition of $\mathcal{G}^p$ and note that $\inf\{J : J \in \hat{\mathcal{J}}\} \geq \underline{J}$ holds wpa1 $\mathcal{H}$-uniformly by Lemma E.5. Recalling the Lebesgue constant $\|\Pi_J\|_\infty$ from Appendix D.3, we may then deduce

$$\|\tilde{h}_J - h_0\|_\infty \geq \|h_{0,J} - h_0\|_\infty \geq (1 + \|\Pi_J\|_\infty)^{-1} \|h_0 - \Pi_J h_0\|_\infty \geq v\underline{B} J^{-p/d} \,,$$

for all $J \in \hat{\mathcal{J}}$ wpa1, uniformly for all $h_0 \in \mathcal{G}^p$ and all $p \in [\underline{p}, \overline{p}]$. It follows by (44) and the discussion preceding the statement of this theorem that

$$\begin{aligned}
\|\hat{h}_J - \hat{h}_{J_0}\|_\infty &\geq v\underline{B} J^{-p/d} - (C_{E.1}+1)\overline{B} J_0^{-p/d} - \|\hat{h}_J - \tilde{h}_J - (\hat{h}_{J_0} - \tilde{h}_{J_0})\|_\infty \\
&\geq (v\underline{B}\beta^{-\underline{p}/d} - (C_{E.1}+1)\overline{B}) J_0^{-p/d} - \|\hat{h}_J - \tilde{h}_J - (\hat{h}_{J_0} - \tilde{h}_{J_0})\|_\infty \\
&> C_2(\xi+1)\tau_{J_0} \frac{\sqrt{J_0}\theta^*_{1-\hat{\alpha}}}{\sqrt{n}} - \|\hat{h}_J - \tilde{h}_J - (\hat{h}_{J_0} - \tilde{h}_{J_0})\|_\infty \,,
\end{aligned}$$

where the second line uses $J < \beta J_0$ and the third uses definition of E and $J_0(p,E)$. It now follows by the preceding display and (51) that

$$\begin{aligned}
&\sup_{p\in[\underline{p},\overline{p}]}\sup_{h_0\in\mathcal{G}^p}\mathbb{P}_{h_0}\big(\hat{J}<\beta J_0(p,E)\big)\\
&\le \sup_{p\in[\underline{p},\overline{p}]}\sup_{h_0\in\mathcal{G}^p}\mathbb{P}_{h_0}\bigg(\inf_{J\in\hat{\mathcal{J}}:J<\beta J_0}\sup_{x\in\mathcal{X}}\frac{\sqrt{n}|\hat{h}_J(x)-\hat{h}_{J_0}(x)|}{\|\hat{\sigma}_{x,J,J_0}\|_{sd}}\le\xi\theta^*_{1-\hat{\alpha}}\bigg)\\
&\le \sup_{p\in[\underline{p},\overline{p}]}\sup_{h_0\in\mathcal{G}^p}\mathbb{P}_{h_0}\bigg(\sup_{(x,J,J_2)\in\hat{\mathcal{S}}}\frac{\sqrt{n}|\hat{h}_J(x)-\hat{h}_{J_2}(x)-(\tilde{h}_J(x)-\tilde{h}_{J_2}(x))|}{\|\hat{\sigma}_{x,J,J_2}\|_{sd}}>\theta^*_{1-\hat{\alpha}}\bigg)+o(1)\\
&\le \sup_{h_0\in\mathcal{H}}\mathbb{P}_{h_0}\bigg(\sup_{(x,J,J_2)\in\hat{\mathcal{S}}}\frac{\sqrt{n}|\hat{h}_J(x)-\hat{h}_{J_2}(x)-(\tilde{h}_J(x)-\tilde{h}_{J_2}(x))|}{\|\hat{\sigma}_{x,J,J_2}\|_{sd}}>\theta^*_{1-\hat{\alpha}}\bigg)+o(1)\to 0\,,
\end{aligned}$$

where the final line is by (49).

Part (i), step 2: Recall $J_0^+(p,D)$ from part (i), step 1 of the proof of Theorem 4.1. By the previous step of this proof and part (i), step 1 of the proof of Theorem 4.1, we have

$$\inf_{p\in[\underline{p},\overline{p}]}\inf_{h_0\in\mathcal{G}^p}\mathbb{P}_{h_0}\big(\beta J_0(p,E)\le\hat{J}\le J_0^+(p,D)\big)\to 1\,. \tag{52}$$

Therefore, by (44), (51), (52), and definition of $\overline{B}$, for every $h_0\in\mathcal{G}^p$ and $x\in\mathcal{X}$ we have

$$\frac{|\tilde{h}_{\hat{J}}(x)-h_0(x)|}{\|\hat{\sigma}_{x,\hat{J}}\|_{sd}}\le(C_{E.1}+1)C_3\overline{B}\frac{\hat{J}^{-p/d}}{\tau_{\hat{J}}\sqrt{\hat{J}}}\le(C_{E.1}+1)C_3\overline{B}\beta^{-\overline{p}/d}2^p\frac{(2^dJ_0(p,E))^{-p/d}}{\tau_{\lceil\beta J_0(p,E)\rceil}\sqrt{\beta J_0(p,E)}}\,,$$

wpa1 uniformly for $h_0\in\mathcal{G}^p$ and $p\in[\underline{p},\overline{p}]$ and $x\in\mathcal{X}$, where $\tau_{\lceil\beta J_0(p,E)\rceil}$ denotes the ill-posedness at resolution level $\inf\{J\in\mathcal{T}:J\ge\beta J_0(p,E)\}$. It now follows from definition of $2^dJ_0(p,E)\equiv J_0^+(p,E)$ from (39) that whenever the preceding inequality holds, we have

$$\sup_{x\in\mathcal{X}}\sqrt{n}\frac{|\tilde{h}_{\hat{J}}(x)-h_0(x)|}{\|\hat{\sigma}_{x,\hat{J}}\|_{sd}}\le C_3(C_{E.1}+1)\overline{B}\beta^{-\overline{p}/d-1/2}2^{\overline{p}+d/2}E^{-1}\frac{\tau_{2^dJ_0(p,E)}}{\tau_{\lceil\beta J_0(p,E)\rceil}}\theta^*_{1-\hat{\alpha}}<A_0\theta^*_{1-\hat{\alpha}}\,,$$

where the final inequality holds uniformly for $h_0\in\mathcal{G}^p$ and $p\in[\underline{p},\overline{p}]$ for a constant $A_0>0$

because $\sup_{J\in\mathcal{T}} \tau_{2^d J}/\tau_{\lceil \beta J\rceil} < \infty$ by virtue of mild ill-posedness. Hence for any $A \geq A_0$,

$$\begin{aligned}
&\inf_{h_0\in\mathcal{G}} \mathbb{P}_{h_0}\left(h_0(x) \in C_n(x,A) \ \ \forall \ \ x \in \mathcal{X}\right) \\
&\geq \inf_{p\in[\underline{p},\overline{p}]} \inf_{h_0\in\mathcal{G}^p} \mathbb{P}_{h_0}\left(\sup_{x\in\mathcal{X}} \sqrt{n}\frac{|\hat{h}_{\hat{J}}(x) - h_0(x)|}{\|\hat{\sigma}_{x,\hat{J}}\|_{sd}} \leq z^*_{1-\alpha} + A\theta^*_{1-\hat{\alpha}}\right) + o(1) \\
&\geq \inf_{p\in[\underline{p},\overline{p}]} \inf_{h_0\in\mathcal{G}^p} \mathbb{P}_{h_0}\left(\sup_{x\in\mathcal{X}} \sqrt{n}\frac{|\hat{h}_{\hat{J}}(x) - \tilde{h}_{\hat{J}}(x)|}{\|\hat{\sigma}_{x,\hat{J}}\|_{sd}} \leq z^*_{1-\alpha}\right) + o(1) \\
&\geq \inf_{p\in[\underline{p},\overline{p}]} \inf_{h_0\in\mathcal{G}^p} \mathbb{P}_{h_0}\left(\sup_{(x,J)\in\mathcal{X}\times\underline{\mathcal{J}}_n} \sqrt{n}\frac{|\hat{h}_{J}(x) - \tilde{h}_{J}(x)|}{\|\hat{\sigma}_{x,J}\|_{sd}} \leq z^*_{1-\alpha}\right) + o(1)\,,
\end{aligned}$$

where the final line is because $\hat{J} \in \underline{\mathcal{J}}_n := \{J \in \mathcal{T} : 0.1(\log \bar{J}_{\max}(R_2))^2 \leq J \leq \bar{J}^-_{\max}(R_1)\}$ with $\bar{J}^-_{\max}(R_1) = \sup\{J \in \mathcal{T} : J < \bar{J}_{\max}(R_1)\}$ and $\hat{\mathcal{J}}_- \supseteq \underline{\mathcal{J}}_n$ both hold wpa1 uniformly for $h_0 \in \mathcal{G}^p$ and $p \in [\underline{p},\overline{p}]$; the former holds by (52) and Lemma E.1 and the latter holds by Lemma E.5 and the fact that $\hat{J} = \tilde{J}$ wpa1 $\mathcal{H}$-uniformly. Let $\underline{z}^*_{1-\alpha}$ denote the $1-\alpha$ quantile of $\sup_{(x,J)\in\mathcal{X}\times\underline{\mathcal{J}}_n} |\mathbb{Z}^*_n(x,J)|$. As $\underline{z}^*_{1-\alpha} \leq z^*_{1-\alpha}$ must hold whenever $\hat{\mathcal{J}} \supseteq \underline{\mathcal{J}}_n$, we therefore have

$$\begin{aligned}
&\inf_{h_0\in\mathcal{G}} \mathbb{P}_{h_0}\left(h_0(x) \in C_n(x,A) \ \ \forall \ \ x \in \mathcal{X}\right) \\
&\geq \inf_{p\in[\underline{p},\overline{p}]} \inf_{h_0\in\mathcal{G}^p} \mathbb{P}_{h_0}\left(\sup_{(x,J)\in\mathcal{X}\times\underline{\mathcal{J}}_n} \sqrt{n}\frac{|\hat{h}_{J}(x) - \tilde{h}_{J}(x)|}{\|\hat{\sigma}_{x,J}\|_{sd}} \leq \underline{z}^*_{1-\alpha}\right) + o(1) = (1-\alpha) + o(1)\,,
\end{aligned}$$

where the last equality follows from Theorem E.2(i) and the definition of $\underline{z}^*_{1-\alpha}$.

Part (ii): By Lemmas E.4, E.6, and E.7 and Assumption 4(i), we have

$$\sup_{x\in\mathcal{X}} |C_n(x,A)| \lesssim (1+A)\tau_{\hat{J}}\sqrt{(\hat{J}\log \bar{J}_{\max})/n}$$

wpa1 $\mathcal{H}$-uniformly. Then by (52) with $J_0 = J_0(p,D)$ and $\bar{A} = 1 + A$, we have that

$$\sup_{x\in\mathcal{X}} |C_n(x,A)| \lesssim \bar{A}\tau_{J_0^+}\sqrt{(J_0^+\log \bar{J}_{\max})/n} \lesssim \bar{A}\tau_{J_0}\sqrt{(J_0\log \bar{J}_{\max})/n} \lesssim \bar{A}\frac{\sqrt{\log \bar{J}_{\max}}}{\theta^*_{1-\hat{\alpha}}} J_0^{-p/d}$$

holds wpa1 uniformly for $h_0 \in \mathcal{G}^p$ and $p \in [\underline{p},\overline{p}]$ and for all $A > 0$, where the second inequality follows from the fact that the model is mildly ill-posed and the third is by definition (39). It follows by Lemma E.6 that there is a constant $C > 0$ (independent of

A) for which

$$\inf_{p\in[\underline{p},\overline{p}]}\inf_{h_0\in\mathcal{G}^p}\mathbb{P}_{h_0}\left(\sup_{x\in\mathcal{X}}|C_n(x,A)|\leq C(1+A)(J_0(p,D))^{-p/d}\right)\to 1\,.$$

The result now follows from part (i), step 2 of the proof of Theorem 4.1, which shows that $\inf_{p\in[\underline{p},\overline{p}]}\inf_{h_0\in\mathcal{H}^p}\left(J_n^*(p,E)\leq J_0(p,D)\right)\to 1$ with $J_n^*(p,E)\asymp(n/\log n)^{d/(2(p+\varsigma)+d)}$. ■

Proof of Theorem 4.3. In some of what follows, we use the fact that the sieve dimensions for CDV wavelet bases are linked via $J^+=2^dJ$ for $J\in\mathcal{T}$. A similar argument (with more complicated notation) applies for B-spline bases.

Part (ii): First note by Lemma E.5 and the fact that $\bar{J}_{\max}(R)=\bar{J}^*_{\max}(R)$ (see (40)) holds for any $R>0$ for all n sufficiently large (see part (ii), step 1 of the proof of Theorem 4.1), we have that $J^*_{\max}(R_1)\leq\hat{J}_{\max}\leq J^*_{\max}(R_2)$ wpa1 $\mathcal{H}$-uniformly.

Recall $M_0(p,R_2)$ from (41). By Lemma E.2, for all $p\in[\underline{p},\overline{p}]$ we have that $M_0(p,R_2)\geq M_0(\overline{p},R_2)\geq 2^{-d}J^*_{\max}(R_2)$ holds for all n sufficiently large. Then by Lemmas E.4, E.6, and E.7 and Assumption 4(i), there exist constants $C,C'>0$ for which

$$\sup_{x\in\mathcal{X}}|C_n(x,A)|\leq C(1+A)\tau_{\tilde{J}}\sqrt{(\tilde{J}\log(\bar{J}^*_{\max}(R_2)))/n}+A\tilde{J}^{-\underline{p}/d}\leq C'(1+A)(J^*_{\max}(R_2))^{-p/d}+A\tilde{J}^{-\underline{p}/d}$$

holds wpa1 uniformly for $h_0\in\mathcal{H}^p$ and $p\in[\underline{p},\overline{p}]$, where the second inequality is by definition of $M_0(p,R_2)$. The proofs of Theorem 4.1 and Corollary 4.1 show that $\bar{J}^*_{\max}(R_2)\asymp(\log n)^{d/\varsigma}$ in the severely ill-posed case. Therefore, it suffices to show that there is a constant $c>0$ for which $\hat{J}\geq c(\log n)^{d/\varsigma}$ holds wpa1 uniformly for $h_0\in\mathcal{G}^p$ and $p\in[\underline{p},\overline{p}]$.

Recall β and E from the proof of Theorem 4.2 and $J_0(p,E)$ from (39). By similar arguments to Lemma E.2, we may deduce that $\inf\{J\in\mathcal{T}:J\geq\beta J_0(p,E)\}>\log\hat{J}_{\max}$ wpa1 uniformly for $h_0\in\mathcal{H}^p$ and $p\in[\underline{p},\overline{p}]$. It then follows by the same argument as part (i), step 1 of the proof of Theorem 4.2 that $\hat{J}\geq\beta J_0(p,E)$ holds wpa1 uniformly for $h_0\in\mathcal{G}^p$ and $p\in[\underline{p},\overline{p}]$. But by Lemma E.6 and the fact that $\log\bar{J}^*_{\max}(R_2)\asymp\log\log n$ for severely ill-posed models, it follows that there is a constant $C''>0$ for which, by defining

$$J^*(p,C'')=\sup\left\{J\in\mathcal{T}:\tau_J\sqrt{(J\log\log n)/n}\leq C''J^{-p/d}\right\},$$

we have $\inf_{p\in[\underline{p},\overline{p}]}\inf_{h_0\in\mathcal{H}^p}\mathbb{P}_{h_0}(J_0(p,E)\geq J^*(p,C''))\to 1$. Finally, we may deduce by a similar argument to part (ii), step 1 of the proof of Theorem 4.1 that $J^*(p,C'')\gtrsim(\log n)^{d/\varsigma}$ for all $p\in[\underline{p},\overline{p}]$, which establishes the desired behavior of $\hat{J}$.

Part (i): By Theorem E.1 and Lemma E.5, there exists a constant $A_0 > 0$ for which

$$|\hat{h}_{\tilde{J}}(x) - h_0(x)| \leq |\hat{h}_{\tilde{J}}(x) - \tilde{h}_{\tilde{J}}(x)| + A_0 \tilde{J}^{-\underline{p}/d}$$

holds for all $x \in \mathcal{X}$ wpa1 $\mathcal{H}$-uniformly. Then for any $A \geq A_0$, we have

$$\inf_{h_0 \in \mathcal{G}} \mathbb{P}_{h_0}\big(h_0(x) \in C_n(x, A) \ \ \forall\, x \in \mathcal{X}\big) \geq \inf_{h_0 \in \mathcal{G}} \mathbb{P}_{h_0}\left(\sup_{x \in \mathcal{X}} \left| \sqrt{n} \frac{\hat{h}_{\tilde{J}}(x) - \tilde{h}_{\tilde{J}}(x)}{\|\hat{\sigma}_{x,\tilde{J}}\|_{sd}} \right| \leq z^*_{1-\alpha} \right) + o(1)\,.$$

Suppose that $J^*_{\max}(R_2) \geq 2^{2d} J^*_{\max}(R_1) \in \mathcal{T}$. Then by definition of $\bar{J}^*_{\max}(R)$ and Remark E.1, we have

$$\frac{\tau_{J^*_{\max}(R_2)}}{\tau_{2^{2d} J^*_{\max}(R_1)}} \asymp \frac{\tau_{J^*_{\max}(R_2)}}{\tau_{2^{2d} J^*_{\max}(R_1)}} \frac{J^*_{\max}(R_2)\sqrt{\log J^*_{\max}(R_2)}}{2^{2d} J^*_{\max}(R_1)\sqrt{\log J^*_{\max}(R_1)}} \leq \frac{R_2}{R_1}\,. \tag{53}$$

But note that if $J^*_{\max}(R_2) \geq 2^{2d} J^*_{\max}(R_1)$ then by severe ill-posedness we have

$$\frac{\tau_{J^*_{\max}(R_2)}}{\tau_{2^d J^*_{\max}(R_1)}} \geq \frac{\tau_{2^{2d} J^*_{\max}(R_1)}}{\tau_{2^d J^*_{\max}(R_1)}} \asymp e^{C((2^{2d} J^*_{\max}(R_1))^{\varsigma/d} - (2^d J^*_{\max}(R_1))^{\varsigma/d})} = e^{C 2^{\varsigma}(2^{\varsigma}-1)(J^*_{\max}(R_1))^{\varsigma/d}} \to +\infty\,,$$

which contradicts (53). Therefore, $\bar{J}^*_{\max}(R_1) \in \{2^{-d}\bar{J}^*_{\max}(R_2), \bar{J}^*_{\max}(R_2)\}$ holds for all n sufficiently large, from which it follows by Lemma E.5 that $\hat{J}_{\max} \in \{2^{-d}\bar{J}^*_{\max}(R_2), \bar{J}^*_{\max}(R_2)\}$ wpa1 $\mathcal{H}$-uniformly. Therefore, $\tilde{J} \leq 2^{-d}\bar{J}^*_{\max}(R_2)$ holds wpa1 $\mathcal{H}$-uniformly. But by part (ii) we also have that $\tilde{J} \geq c\bar{J}^*_{\max}(R_2)$ holds wpa1 uniformly $h_0 \in \mathcal{G}^p$ and $p \in [\underline{p}, \overline{p}]$. Therefore, $\tilde{J} \in \underline{\mathcal{J}}_n := \{J \in \mathcal{T} : c\bar{J}^*_{\max}(R_2) \leq J \leq 2^{-d}\bar{J}_{\max}(R_2)\}$ and $\hat{\mathcal{J}} \supseteq \underline{\mathcal{J}}_n$ both hold wpa1 uniformly for $h_0 \in \mathcal{G}^p$ and $p \in [\underline{p}, \overline{p}]$.

Let $\underline{z}^*_{1-\alpha}$ denote the $1-\alpha$ quantile of $\sup_{(x,J)\in\mathcal{X}\times\underline{\mathcal{J}}_n} |\mathbb{Z}^*_n(x, J)|$. As $\underline{z}^*_{1-\alpha} \leq z^*_{1-\alpha}$ must hold whenever $\hat{\mathcal{J}} \supseteq \underline{\mathcal{J}}_n$, we therefore have

$$\begin{aligned}
&\inf_{h_0 \in \mathcal{G}} \mathbb{P}_{h_0}\left(h_0(x) \in C_n(x, A) \ \ \forall \ \ x \in \mathcal{X}\right) \\
&\geq \inf_{p \in [\underline{p}, \overline{p}]} \inf_{h_0 \in \mathcal{G}^p} \mathbb{P}_{h_0}\left(\sup_{(x,J)\in\mathcal{X}\times\underline{\mathcal{J}}_n} \sqrt{n} \frac{|\hat{h}_J(x) - \tilde{h}_J(x)|}{\|\hat{\sigma}_{x,J}\|_{sd}} \leq \underline{z}^*_{1-\alpha} \right) + o(1) = (1-\alpha) + o(1)\,,
\end{aligned}$$

where the last equality follows from Theorem E.2(i) and the definition of $\underline{z}^*_{1-\alpha}$. ■

E.4 Supplemental Results: UCBs for Derivatives

Here we present supplemental results for the proofs of Theorems 4.4 and 4.5. Throughout this subsection, for any fixed $R > 0$, let $\bar{J}_{\max} \equiv \bar{J}_{\max}(R)$. Also let $J_{\min} \to \infty$ as $n \to \infty$ with $J_{\min} \leq \bar{J}_{\max}$. Define $\mathcal{J}_n = \{J \in \mathcal{T} : J_{\min} \leq J \leq \bar{J}_{\max}\}$. Also recall δ_n from (43). We introduce the bootstrap process for the derivatives:

$$\mathbb{Z}_n^{a*}(x, J) \equiv \frac{D_J^{a*}(x)}{\hat{\sigma}_J^a(x)} = \frac{1}{\|\hat{\sigma}_{x,J}^a\|_{sd}} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{L}_{J,x}^a b_{W_i}^{K(J)} \hat{u}_{i,J} \varpi_i \right),$$

where $\|\hat{\sigma}_{x,J}^a\|_{sd}^2 \equiv n\hat{\sigma}_J^{a2}(x) = \hat{L}_{J,x}^a \widehat{\Omega}_{J,J} (\hat{L}_{J,x}^a)'$ and $\hat{L}_{J,x}^a = (\partial^a \psi_x^J)' [\widehat{S}_J' \widehat{G}_{b,J}^{-1} \widehat{S}_J]^{-1} \widehat{S}_J' \widehat{G}_{b,J}^{-1}$ with $\partial^a \psi_x^J$ denoting the derivative applied element-wise: $\partial^a \psi_x^J = (\partial^a \psi_{J1}(x), \ldots, \partial^a \psi_{JJ}(x))'$. Proofs of these supplemental results are presented in our earlier working paper version Chen et al. (2022), where they are labelled as Lemmas E.12, E.13, and E.14, respectively.

Lemma E.8 *Let Assumptions 1-3 hold. Then: there is a universal constant $C_{E.8} > 0$ such that*

$$\inf_{h_0 \in \mathcal{H}} \mathbb{P}_{h_0} \left(\sup_{(x,J) \in \mathcal{X} \times \mathcal{J}_n} \left| \frac{\|\hat{\sigma}_{x,J}^a\|_{sd}^2}{\|\sigma_{x,J}^a\|_{sd}^2} - 1 \right| \leq C_{E.4} \delta_n \right) \to 1 .$$

Lemma E.9 *Let Assumptions 1-4 hold. For a given $\alpha \in (0, 1)$, let $z_{1-\alpha}^{a*}$ denote the $1-\alpha$ quantile of $\sup_{(x,J) \in \mathcal{X} \times \hat{\mathcal{J}}} |\mathbb{Z}_n^{a*}(x, J)|$. Then: with $\bar{J}_{\max}(R)$ as defined in (38) for any $R > 0$, there exists a constant $C_{E.9} > 0$ for which*

$$\inf_{h_0 \in \mathcal{H}} \mathbb{P}_{h_0} \left(z_{1-\alpha}^{a*} \leq C_{E.9} \sqrt{\log \bar{J}_{\max}(R)} \right) \to 1 .$$

Lemma E.10 *Let Assumptions 1-4 hold and let $J_{\min} \asymp (\log \bar{J}_{\max})^2$. Then: there exists a sequence $\gamma_n \downarrow 0$ for which*

$$\sup_{s \in \mathbb{R}} \left| \mathbb{P}_{h_0} \left(\sup_{(x,J) \in \mathcal{X} \times \mathcal{J}_n} \left| \sqrt{n} \frac{\partial^a \hat{h}_J(x) - \partial^a \tilde{h}_J(x)}{\|\hat{\sigma}_{x,J}^a\|_{sd}} \right| \leq s \right) - \mathbb{P}^* \left(\sup_{(x,J) \in \mathcal{X} \times \mathcal{J}_n} |\mathbb{Z}_n^{a*}(x, J)| \leq s \right) \right| \leq \gamma_n$$

holds wpa1 $\mathcal{H}$-uniformly.

E.5 Proofs of Theorems 4.4 and 4.5 on UCBs for Derivatives

Proof of Theorem 4.4. The proof follows similar arguments to the proof of Theorem 4.2. Here we state the necessary modifications.

Part (i), step 1: Identical to part (i), step 1 of the proof of Theorem 4.2.

Part (i), step 2: Note that by Theorem E.1 and a similar argument to the proof of Corollary 3.1 of Chen and Christensen (2018), we have

$$\inf_{h_0\in\mathcal{H}} \mathbb{P}_{h_0}\left(\|\partial^a \tilde{h}_J - \partial^a h_0\|_\infty \leq C_6 J^{(|a|-p)/d} \quad \forall\, J \in [1, \bar{J}_{\max}] \cap \mathcal{T} \right) \to 1$$

for some constant $C_6 > 0$. Moreover, by Lemma E.8 and Assumption 4(iii) there is a constant $C_7 > 0$ for which

$$\inf_{h_0\in\mathcal{H}} \mathbb{P}_{h_0}\left(\sup_{(x,J)\in\mathcal{X}\times\hat{\mathcal{J}}} \frac{\tau_J J^{1/2+|a|/d}}{\|\hat{\sigma}^a_{x,J}\|_{sd}} \leq C_7 \right) \to 1\,.$$

It now follows by (52) that

$$\frac{|\partial^a \tilde{h}_{\hat{J}}(x) - \partial^a h_0(x)|}{\|\hat{\sigma}^a_{x,\hat{J}}\|_{sd}} \leq C_6 C_7 \frac{\hat{J}^{-p/d}}{\tau_{\hat{J}}\sqrt{\hat{J}}} \leq C_6 C_7 \beta^{-\overline{p}/d} 2^{\overline{p}} \frac{(2^d J_0(p,E))^{-p/d}}{\tau_{\lceil \beta J_0(p,E)\rceil}\sqrt{\beta J_0(p,E)}}\,,$$

wpa1 uniformly for $h_0 \in \mathcal{G}^p$ and $p \in [\underline{p}, \overline{p}]$ and $x \in \mathcal{X}$. The remainder of the proof of this part now follows by identical arguments to part (i), step 2 of the proof of Theorem 4.2, using Lemma E.10 in place of Theorem E.2(i).

Part (ii): By Lemma E.6, Lemmas E.8 and E.9 and Assumption 4(iii), we have

$$\sup_{x\in\mathcal{X}} |C^a_n(x,A)| \lesssim (1+A)\tau_{\hat{J}} \hat{J}^{1/2+|a|/d}\sqrt{(\log \bar{J}_{\max})/n}$$

wpa1 $\mathcal{H}$-uniformly. Then by display (52), with $J_0 = J_0(p,D)$ we have that

$$\begin{aligned}\sup_{x\in\mathcal{X}} |C^a_n(x,A)| &\lesssim (1+A)\tau_{J_0^+}(J_0^+)^{1/2+|a|/d}\sqrt{(\log \bar{J}_{\max})/n} \\ &\lesssim (1+A)\tau_{J_0} J_0^{|a|/d}\sqrt{(J_0 \log \bar{J}_{\max})/n} \lesssim (1+A)\frac{\sqrt{\log \bar{J}_{\max}}}{\theta^*_{1-\hat{\alpha}}} J_0^{(|a|-p)/d}\end{aligned}$$

holds wpa1 uniformly for $h_0 \in \mathcal{G}^p$ and $p \in [\underline{p}, \overline{p}]$, where the second inequality follows from the fact that the model is mildly ill-posed and the third is by definition (39). The result now follows by similar arguments to part (ii) of the proof of Theorem 4.2. ■

Proof of Theorem 4.5. The proof follows similar arguments to the proof of Theorem 4.3. Here we state the necessary modifications.

Part (i): By Lemma E.5, Theorem E.1, and similar arguments to the proof of Corollary

3.1 of Chen and Christensen (2018), there exists a constant $A_0 > 0$ for which

$$|\partial^a \hat{h}_{\tilde{J}}(x) - \partial^a h_0(x)| \leq |\partial^a \hat{h}_{\tilde{J}}(x) - \partial^a \tilde{h}_{\tilde{J}}(x)| + A_0 \tilde{J}^{(|a|-\underline{p})/d}$$

holds for all $x \in \mathcal{X}$ wpa1 $\mathcal{H}$-uniformly. Then for any $A \geq A_0$, we have

$$\inf_{h_0 \in \mathcal{G}} \mathbb{P}_{h_0}\big(\partial^a h_0(x) \in C_n^a(x, A) \ \ \forall\, x \in \mathcal{X}\big) \geq \inf_{h_0 \in \mathcal{G}} \mathbb{P}_{h_0}\left(\sup_{x \in \mathcal{X}} \left| \sqrt{n} \frac{\partial^a \hat{h}_{\tilde{J}}(x) - \partial^a \tilde{h}_{\tilde{J}}(x)}{\|\hat{\sigma}^a_{x,\tilde{J}}\|_{sd}} \right| \leq z^{a*}_{1-\alpha} \right) + o(1)\,.$$

The remainder of the proof now follows similarly to the proof of Theorem 4.3, using Lemma E.10 in place of Theorem E.2(i).

Part (ii): By Lemmas E.2, E.6, E.8, and E.9 and Assumption 4(iii), there exist constants $C, C' > 0$ for which

$$\begin{aligned} \sup_{x \in \mathcal{X}} |C_n^a(x, A)| &\leq C(1 + A)\tau_{\tilde{J}} \tilde{J}^{1/2+|a|/d} \sqrt{\log(\bar{J}^*_{\max}(R_2))/n} + A\tilde{J}^{(|a|-\underline{p})/d} \\ &\leq C'(1 + A)(J^*_{\max}(R_2))^{(|a|-p)/d} + A\tilde{J}^{(|a|-\underline{p})/d} \end{aligned}$$

holds wpa1 uniformly for $h_0 \in \mathcal{H}^p$ and $p \in [\underline{p}, \overline{p}]$. The remainder of the proof now follows similarly to the proof of Theorem 4.3. ■

References

Beare, B. K. (2010). Copulas and temporal dependence. *Econometrica 78*(1), 395–410.

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

Chen, X., T. Christensen, and S. Kankanala (2022). Adaptive estimation and uniform confidence bands for nonparametric structural functions and elasticities. arxiv:2107.11869v2 [econ.em], (dated October 11, 2022).

Chen, X. and T. M. Christensen (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics 188*(2), 447–465.

Chen, X. and T. M. Christensen (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression. *Quantitative Economics 9*(1), 39–84.

Cohen, A., I. Daubechies, and P. Vial (1993). Wavelets on the interval and fast wavelet transforms.

DeVore, R. A. and G. G. Lorentz (1993). *Constructive approximation*. Springer.

DeVore, R. A. and V. A. Popov (1988). Interpolation of Besov spaces. *Transactions of the American Mathematical Society 305*(1), 397–414.

Giné, E. and R. Nickl (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press.

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