ADAPTIVE, RATE-OPTIMAL TESTING IN INSTRUMENTAL VARIABLES MODELS

By

Christoph Breunig and Xiaohong Chen

June 2020

COWLES FOUNDATION DISCUSSION PAPER NO. 2238

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

http://cowles.yale.edu/

Adaptive, Rate-Optimal Testing in Instrumental Variables Models*

Christoph Breunig[†] Xiaohong Chen[‡]

First version: August 2018, Revised June 16, 2020

This paper proposes simple, data-driven, optimal rate-adaptive inferences on a structural function in semi-nonparametric conditional moment restrictions. We consider two types of hypothesis tests based on leave-one-out sieve estimators. A structure-space test (ST) uses a quadratic distance between the structural functions of endogenous variables; while an image-space test (IT) uses a quadratic distance of the conditional moment from zero. For both tests, we analyze their respective classes of nonparametric alternative models that are separated from the null hypothesis by the *minimax rate of testing*. That is, the sum of the type I and the type II errors of the test, uniformly over the class of nonparametric alternative models, cannot be improved by any other test. Our new minimax rate of ST differs from the known minimax rate of estimation in nonparametric instrumental variables (NPIV) models. We propose computationally simple and novel exponential scan data-driven choices of sieve regularization parameters and adjusted chi-squared critical values. The resulting tests attain the minimax rate of testing, and hence optimally adapt to the unknown smoothness of functions and are robust to the unknown degree of ill-posedness (endogeneity). Data-driven confidence sets are easily obtained by inverting the adaptive ST. Monte Carlo studies demonstrate that our adaptive ST has good size and power properties in finite samples for testing monotonicity or equality restrictions in NPIV models. Empirical applications to nonparametric multi-product demands with endogenous prices are presented.

Keywords: Instrumental variables; Minimax rate of testing; Adaptive testing; Exponential scan; Confidence sets; Quadratic functionals; Shape restrictions.

*We thank Don Andrews, Tim Armstrong, Denis Chetverikov, Tim Christensen, Enno Mammen, Peter Mathe and others at numerous workshops and conferences for helpful comments. The empirical work here is researchers own analyses based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researchers and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

[†]Department of Economics, Emory University, Rich Memorial Building, Atlanta, GA 30322, USA. Email: christoph.breunig@emory.edu

[‡]Cowles Foundation for Research in Economics, Yale University, Box 208281, New Haven, CT 06520, USA. Email: xiaohong.chen@yale.edu

1. Introduction

Models with endogeneity are pervasive in economics, and are one of the most defying features that differentiate econometrics from statistics. In the big data era, semiparametric and nonparametric methods and models allowing for flexible endogeneity are increasingly widely used in empirical research. A common difficulty in applying any semiparametric and nonparametric methods in practice is how to choose tuning (regularization) parameters in a simple, data-driven way that still possesses some "optimal" theoretical properties. For nonparametric models with endogeneity such as nonparametric instrumental variables (NPIV) and nonparametric quantile instrumental variables (NPQIV), it is well-known that the finite sample performance of various estimators and tests are much more sensitive to tuning parameters than those without endogeneity.

There are a few papers on data-driven choices of regularization parameters in estimation of a NPIV model.[1] However, it is well known that data-driven choices designed for "optimal" nonparametric estimation do not lead to "optimal" inference (testing and confidence sets) in nonparametric settings (see, e.g., Gine and Nickl [2016]). To the best of our knowledge, there is currently no work on minimax rate-optimal testing for NPIV models, nor on data-driven choice of regularization parameters for rate-optimal testing and confidence sets on NPIV models. In this paper we shall address this important issue within the framework of rate-optimal testing in semiparametric and nonparametric conditional moment restrictions. As a leading example, we shall provide computationally simple, data-driven choices of tuning parameters for optimal inference on NPIV functions such as multi-product demand functions (of endogenous prices) in industrial organization.

This paper first considers the minimax rate-optimal hypothesis testing in semiparametric or nonparametric models defined by conditional moment restrictions. The maintained modeling assumption is that there is a nonparametric structural function h satisfying

$$\mathrm{E}[\rho(Y, h(X))|W] = 0, \tag{1.1}$$

where ρ is a possibly non-smooth mapping that is known up to the function h, X is a d_x-dimensional vector of continuous endogenous regressors, W is a d_w-dimensional vector of conditional (instrumental) variables, and the joint distribution of (Y, X, W) is unspecified. Our goal is to test whether h coincides with a restricted function h^{R}, such as parametric or semiparametric or shape-restricted function. For example, in a NPIV model $\mathrm{E}[Y - h(X)|W] = 0$ we would be interested in testing whether the structural NPIV function h coincides with some decreasing function h^{R}. We propose two statistics for the hypothesis

[1] See, e.g., Horowitz [2014], Liu and Tao [2014], Centorrino [2014], Chen and Christensen [2015], Breunig and Johannes [2016], Gautier and Le Pennec [2018], Jansson and Pouzo [2019] and the references therein. Most papers suggest data-driven procedures without establishing the rate-adaptivity in NPIV estimation.

under consideration: a structure-space test (ST) and an image-space test (IT). The ST is based on a squared distance between a nonparametric estimator of h and its restricted version h^{R}. The IT is based on the squared distance between a nonparametric estimator of $\mathrm{E}[\rho(Y, h^{\mathrm{R}}(X))|W]$ and zero.

Both test statistics are based on simple leave-one-out sieve estimators of some quadratic functionals. We establish an upper bound on the sum of the type I and type II errors. Specifically, we bound the type I error uniformly over distributions satisfying the null hypothesis and the type II error uniformly over a class of nonparametric alternative models separated from the null hypothesis via a so-called *rate of testing*. We then establish a lower bound for the sum of the type I and the type II errors at the same separation rate. Thus, there exists no other test that provides a better performance with respect to the sum of those errors. This optimal rate of separation is called the *minimax rate of testing*.

A key technical part to establish the minimax rate of our ST in quadratic distance is to derive a tight upper bound on the convergence rate of a leave-one-out sieve estimator of a quadratic functional of a NPIV function h; see the online Appendix F.[2] This rate is different from the existing minimax rate of convergence in root-mean squared error of any consistent estimator to the function h itself. However, the minimax rate of our ST depends on an optimal choice of sieve dimension (the key tuning parameter), which is determined by the unknown degree of ill-posedness (due to the endogeneity in (1.1)) and the unknown regularity of the nonparametric alternative functions different from the null restricted functions.

We propose a computationally simple, data-driven version of our ST that does not require a priori knowledge of smoothness of nonparametric alternative functions h nor the degree of ill-posedness. The data-driven test rejects the null hypothesis as soon as there is a sieve dimension (say the smallest sieve dimension) in an admissible index set such that the corresponding normalized quadratic distance estimator exceeds one; and fails to reject the null when the maximal (over the admissible index set) normalized quadratic distance estimator is less than or equal to one. The cardinality of this admissible index set is determined by a novel exponential scan (ES) method that automatically takes the unknown degree of ill-posedness (endogeneity) into account. We show that our data-driven ST attains the minimax optimal rate for severely ill-posed problems and is within a $\log\log(n)$ term for mildly ill-posed problems, where n is the sample size. This extra $\log\log(n)$ term is the price to pay for adaptivity to unknown smoothness of the nonparametric alternative functions different from the null restricted function h^{R}.

By inverting the adaptive tests we obtain confidence sets on restricted (constrained)

[2]We prove in Appendix F that this convergence rate coincides with the lower bound derived by Chen and Christensen [2018] for estimation of a quadratic functional of a NPIV function. As shown in Chen and Christensen [2018], the plug-in sieve estimator does not achieve the optimal minimax rate for estimation of a quadratic functional of a NPIV in the mildly-illposed case.

structural functions h^{R}. These confidence sets do not require additional choices of tuning parameters. The adaptive minimax rate of testing determines the radius of the confidence sets. We argue that the radius based on our adaptive ST can only be marginally improved in a very limited range of submodels depending on the regularity of the unknown function h in NPIV models.

Monte Carlo studies indicate that our data-driven ST is not only computationally very fast, but also has accurate size and good power in finite samples, without the need of computationally intensive bootstrap critical values. In a simulation study of hypothesis testing for monotonicity of a NPIV function, our adaptive ST automatically leads to a data-driven confidence set under monotonicity restrictions if the null is not rejected. When the null of monotonicity is rejected by our adaptive ST, the data-driven choice of the smallest sieve dimension leading to the null rejection can still lead to a consistent sieve estimate of the unrestricted NPIV function h, while parts of the true h and its sieve estimate lie outside of the monotonicity-constrained confidence sets. This simulation demonstrates the importance of data-driven choice of tuning parameters for testing shapes of a NPIV function. We provide empirical applications concerning shapes of consumer demands, where our data-driven test detects heterogeneity in the curvature of demand curves among different income groups. For instance, our adaptive ST fails to reject that the demand for certain nondurable goods is decreasing (in its price) for low income household, but does reject the decreasing shape for high income household. Therefore, it may lead to erroneous policy evaluations when nonparametric decreasing demand (in own price) is imposed across all income levels.

Our main contribution is the data-driven, rate-optimal hypothesis testing in structure-space. But we also present the minimax rate-adaptive image-space test (IT) as comparison. Although both are simple to implement, their data-driven procedures choose different key tuning parameters to achieve their respective minimax optimal rate for testing. The sieve dimension J for approximating $h(X)$ is the key tuning parameter in the ST approach, while the sieve dimension K for approximating the conditional moment function $\mathrm{E}[\rho(Y, h^{\mathrm{R}}(X))|W]$ is the key tuning parameter in the IT approach for a simple or parametric null hypothesis. The adaptive ST has the advantage of automatically providing a data-driven choice of the sieve dimension J that can be used for estimation of semi-nonparametric or shape-restricted function h^{R}. This greatly simplifies the construction of data-driven confidence sets. In addition, we show both theoretically and via Monte Carlo simulations that the adaptive ST can be more powerful than adaptive IT when the dimension of the conditional instruments is larger than the dimension of endogenous regressors (i.e., $d_w > d_x$). On the other hand, the image-space test (IT) is more convenient for non-separable models such as nonparametric quantile IV regressions, as well as for partially-identified models.

Literature review: The concept of minimax rate of testing in nonparametric models was perhaps first introduced by Ingster [1993] and Spokoiny [1996].[3] It has been applied to optimal testing in nonparametric regression models without endogeneity, including Horowitz and Spokoiny [2001] and Guerre and Lavergne [2005] and others. Our paper is the first to study minimax rate-optimal test in nonparametric conditional moment restrictions with endogeneity, including NPIV model as a leading example.

There are papers on specification tests for NPIV type models by extending Bierens [1990]'s test for conditional moment restrictions to models that allow for functions depending on endogenous regressors; see, e.g., Horowitz [2006], Breunig [2015], Santos [2012], Tao [2014], Chernozhukov et al. [2015], Zhu [2020] and the references therein. These tests are similar to what we called the image-space test. Among these papers, Chernozhukov et al. [2015] is the most general one that provides inference on equality or/and inequality constrained conditional moment restrictions allowing for partial identification. Chen and Pouzo [2015] provide inference results using either sieve Wald ("structural space") test or sieve QLR ("image space") tests for general point-identified semi-nonparametric conditional moment restriction models. Chetverikov and Wilhelm [2017] studied mean squared rate sieve estimation of NPIV by imposing monotonicity restriction. Freyberger and Reeves [2019] considered L^2 confidence sets for monotone NPIV function. Compiani [2019] also imposed monotone restriction in his estimation of IO demand NPIV function. None of these papers consider minimax tests for NPIV type models nor data-driven choice of key tuning parameters. Our paper is the first to propose simple, adaptive structure- and image- space tests that achieves minimax rate-optimality. In addition, we provide data-driven choice of tuning parameters based confidence sets for NPIV functions h^{R}.

The remainder of the paper is organized as follows. Section 2 describes our data-driven structure-space hypothesis testing (ST). It also presents a simulation study to adaptive testing for monotonicity in NPIV models. Section 3 first establishes the minimax optimal rate of the ST. It then shows that this optimal rate is attained (within a $\log\log n$ term) by our data-driven ST procedure. Section 4 introduces the data-driven image-space test statistic, and presents its minimax rate of testing and the adaptivity. Section 5 provides three empirical illustrations. It also contains additional Monte Carlo studies to compare the finite-sample size and power properties of the adaptive ST vs the adaptive IT. Section 6 briefly concludes. Appendices A and B contain proofs for the minimax rates for the ST and the adaptive ST under simple null hypothesis respectively. The online supplementary appendices contain additional materials: Appendix C presents robustness checks using bootstrap critical values for the empirical applications. Appendices D and E provide proofs for the optimal rates of the adaptive ST under composite null hypothesis and of the adaptive IT respectively. Appendices F and G contain additional technical lemmas.

[3] It has a close connection to robustness or sensitivity literature that has gain popularity in macroeconomics; see, e.g., Hansen and Sargent [2008].

2. Preview of Adaptive Structure-Space Test

We first introduce the null and the alternative hypotheses as well as the concept of minimax rate of testing in Subsection 2.1. We then describe our new data-driven, rate-adaptive structure-space test (ST) for NPIV type models in Subsection 2.2. The formal theoretical justifications are postponed to Section 3. Subsection 2.3 provides a simulation study of our adaptive testing for monotonicity of structural functions.

2.1. The Hypotheses for NPIV models

Let $\mathcal{H}$ denote some class of functions. Let $\{(Y_i, X_i, W_i)\}_{i=1}^n$ be a random sample from the distribution P_h of (Y, X, W), where $h \in \mathcal{H}$, such that

$$\mathrm{E}[Y - h(X)|W] = 0.$$

Let $\mathcal{H}^{\mathrm{R}}$ denote a subset of functions in $\mathcal{H}$ that satisfies a conjectured restriction, such as monotonicity, concavity, some other shape or some parametric restrictions. For any $h \in \mathcal{H}$, we introduce $h^{\mathrm{R}} \in \mathcal{H}^{\mathrm{R}}$ such that $\mathrm{E}\,|\,\mathrm{E}[h(X) - h^{\mathrm{R}}(X)|W]|^2 \leq \mathrm{E}\,|\,\mathrm{E}[h(X) - h^{r}(X)|W]|^2$ for all $h^r \in \mathcal{H}^{\mathrm{R}}$.

We analyze the null hypothesis that there exists a function $h \in \mathcal{H}$ with $\mathrm{E}[Y - h(X)|W] = 0$ satisfying a conjectured restriction captured by $\mathcal{H}^{\mathrm{R}}$, specifically, the set

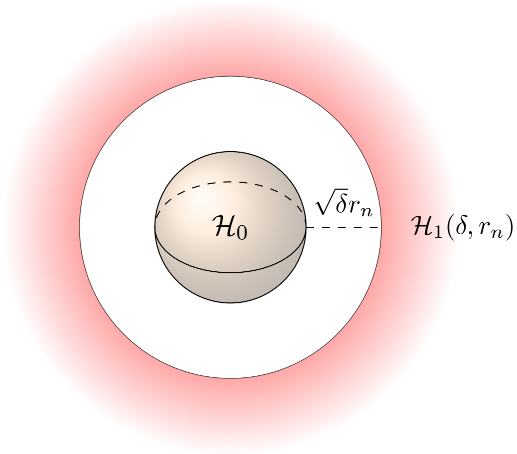
$$\mathcal{H}_0 := \Big\{ h \in \mathcal{H} :\ \mathrm{E}[Y - h(X)|W] = 0 \quad \text{and} \quad \int \big(h(x) - h^{\mathrm{R}}(x)\big)^2 \mu(x) dx = 0 \Big\}$$

is not empty. Here we measure the distance between restricted and unrestricted functions with a measure depending on a prespecified weighting function μ which is restricted to be positive on the support of X. If we want to test that h coincides with h^{R} on some subset of of the support of X only then this modified null hypothesis can be implemented by changing μ accordingly. To analyze the power of any test against nonparametric alternatives, we require some separation between the null and the class of nonparametric alternatives for all $h \in \mathcal{H}$. The resulting class of alternatives considered in this paper is given by

$$\mathcal{H}_1(\delta, r_n) := \Big\{ h \in \mathcal{H} :\ \mathrm{E}[Y - h(X)|W] = 0 \quad \text{and} \quad \int (h(x) - h^{\mathrm{R}}(x))^2 \mu(x) dx \geq \delta\, r_n^2 \Big\}$$

for some constant $\delta > 0$ and a separation rate r_n. The rate r_n is also known as the *rate of testing* and we establish its optimality in the minimax sense as described below.

In this paper, we establish the *minimax rate of testing* r_n in the sense of Ingster [1993]: We propose a test which minimizes the sum of Type I error uniformly over $\mathcal{H}_0$ and the maximum Type II error uniformly over $\mathcal{H}_1(\delta, r_n)$. Moreover, we show that the sum of both errors cannot be improved by any other test.



The minimax rate of testing requires on optimal choice of tuning parameters depending on unknown smoothness properties of the structural function h and unknown mapping properties of the conditional expectation given W. We provide a data driven extension to the minimax test, i.e., a testing procedure that adapts to the smoothness of the unrestricted function h in the presence of unknown smoothing properties of the conditional expectation mapping.

2.2. An adaptive structure-space test for NPIV models

In this section we describe our adaptive structure-space test for the point identified NPIV models. Our test builds on the leave-one-out, series estimator of the quadratic distance $\int (h(x) - h^{\mathrm{R}}(x))^2 \mu(x) dx$ depending on the dimension J given by

$$\widehat{D}_J(h^{\mathrm{R}}) = \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \big(Y_i - h^{\mathrm{R}}(X_i)\big)\big(Y_{i'} - h^{\mathrm{R}}(X_{i'})\big) b^{K(J)}(W_i)' \widehat{A}' G \widehat{A}\, b^{K(J)}(W_{i'})$$

where $G = \int \psi^J(x)\psi^J(x)'\mu(x)dx$ and $\widehat{A} = n[\Psi' B(B'B)^- B'\Psi]^- \Psi' B(B'B)^-$ with $-$ denoting the generalized inverse. Here, $\Psi = (\psi^J(X_1), \ldots, \psi^J(X_n))'$ and $B = (b^K(W_1), \ldots, b^K(W_n))'$ where $\psi^J(\cdot)$ and $b^K(\cdot)$ are vectors of basis functions of dimension J and $K = K(J)$. The dimension parameters J and $K = K(J)$ can grow slowly with the sample size n. Throughout the paper, we keep the relationship of J and K fixed, i.e., the function $K(\cdot)$ does not depend on the sample size.

In many situations, the functions in $\mathcal{H}^{\mathrm{R}}$ are unknown. We consider the restricted, sieve minimum distance estimator of h^{R} given by

$$\widehat{h}_J^{\mathrm{R}} = \operatorname*{arg\,min}_{h \in \mathcal{H}_J^{\mathrm{R}}} \sum_{1 \le i, i' \le n} \big(Y_i - h(X_i)\big)\big(Y_{i'} - h(X_{i'})\big) b^{K(J)}(W_i)'(B'B/n)^- b^{K(J)}(W_{i'}) \qquad (2.1)$$

for some sieve space $\mathcal{H}_J^{\mathrm{R}} = \{\phi = \beta'\psi^J : \phi \in \mathcal{H}^{\mathrm{R}}\}$ which is a subset of $\mathcal{H}^{\mathrm{R}}$ and becomes dense as J tends to infinity.

We propose a test which rejects the null hypothesis $\mathcal{H}_0 \neq \emptyset$, as soon as at least for one J the normalized estimator $\widehat{D}_J(\widehat{h}_J^{\mathrm{R}})$ is sufficiently large. That is, given a nominal level $\alpha \in (0,1)$ we define

$$\widehat{\mathtt{ST}}_n = \mathbb{1}\left\{\exists J \in \widehat{\mathcal{I}}_n \quad \text{such that} \quad \widehat{\mathcal{W}}_J := \frac{n\,\widehat{D}_J(\widehat{h}_J^{\mathrm{R}})}{\sqrt{2}\widehat{\eta}_J(\alpha)\,\widehat{\mathrm{v}}_J(\widehat{h}_J^{\mathrm{R}})} > 1\right\}, \tag{2.2}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function and $\widehat{\mathrm{v}}_J$, $\widehat{\mathcal{I}}_n$ and $\widehat{\eta}_J(\alpha)$ are defined as follows. First,

$$\widehat{\mathrm{v}}_J(h) = \left\| n^{-1}\sum_{i=1}^{n}\big(Y_i - h(X_i)\big)^2 G^{1/2}\widehat{A}\, b^{K(J)}(W_i) b^{K(J)}(W_i)'\widehat{A}' G^{1/2}\right\|_F \tag{2.3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.[4]

Second, the index set $\widehat{\mathcal{I}}_n$ is constructed via a simple *exponential scan* (ES) procedure

$$\widehat{\mathcal{I}}_n = \left\{J \leq \widehat{J}_{\max}: \quad J = \underline{J}2^j \text{ where } j = 0, 1, \ldots, j_{\max}\right\} \tag{2.4}$$

where $\underline{J} := \lfloor 3\sqrt{\log\log n}\rfloor$, $j_{\max} := \lceil \log_2(n^{1/3}/\underline{J})\rceil$, and

$$\widehat{J}_{\max} = \min\left\{J > \underline{J} :\, \zeta^2(J)\sqrt{\ell(J)(\log n)/n} \geq s_{\min}\big((B'B/n)^{-1/2}(B'\Psi/n)G^{-1/2}\big)\right\}$$

where $\ell(J) = 0.1\log\log J$, and $s_{\min}(\cdot)$ is the minimal singular value. Further $\zeta(J) = \sqrt{J}$ for spline, wavelet, or trigonometric sieve basis, and $\zeta(J) = J$ for orthogonal polynomial basis.

Finally, the critical value $\widehat{\eta}_J(\alpha)$ is specified differently for testing equality and inequality constraints. For testing equality constraints, we compute $\widehat{\eta}_J(\alpha)$ using the Bonferroni correction to a critical value from a centralized chi-square distribution relative to the cardinality of the ES index set denoted by $\#(\widehat{\mathcal{I}}_n)$. That is,

$$\widehat{\eta}_J(\alpha) = \frac{q\big(\alpha/\#(\widehat{\mathcal{I}}_n), J\big) - J}{\sqrt{2J}}, \tag{2.5}$$

where $q\big(a, J\big)$ denotes the upper a-quantile of χ^2 distribution with J degrees of freedom. For testing inequality constraints we have implemented two approaches. The first one is presented in Remark 3.1, which is a simple, data-driven, finite-dimensional correction to the chi-square critical values. The second one is introduced in Remark 3.2, which is a bootstrap approach to calculate the critical values.

[4] The Frobenius norm for a $J \times J$ matrix $M = (M_{jl})_{1\leq j,l\leq J}$ is defined as $\|M\|_F = \sqrt{\sum_{j,l=1}^{J} M_{jl}^2}$.

2.3. A Monte Carlo Study: Adaptive Testing for Monotonicity

We investigate finite sample performance of our adaptive ST for decreasing function in a Monte Carlo experiments. The results are based on 5000 Monte Carlo replications for every experiment. In all the experiments, realizations of Y are generated according to a NPIV model

$$Y = h(X) + U \ , \quad \mathrm{E}[U|W] = 0 \tag{2.6}$$

for some unknown structural function h. The functional form of h and the joint distribution of (X, W, U) vary in different Monte Carlo designs. Results presented in this section indicate that our adaptive ST with simple data-driven critical values has very good size and power in finite samples, and that the adaptive ST with computationally demanding bootstrapped critical values has no obvious improvement in terms of size or power.

Let Φ denote the standard normal distribution function. We set $X_i = \Phi(X_i^*)$ and $W_i = \Phi(W_i^*)$, where the random vector (X_i^*, W_i^*, U_i) is generated according to

$$\begin{pmatrix} X_i^* \\ W_i^* \\ U_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \xi & 0.3 \\ \xi & 1 & 0 \\ 0.3 & 0 & 1 \end{pmatrix} \right) . \tag{2.7}$$

The parameter ξ captures the strength of instruments and varies in the experiments below. As ξ increases the instrument becomes stronger (or the ill-posedness gets weaker). The experimental design with $\xi = 0.5$ coincides with the one considered by Chernozhukov et al. [2015], and we follow their design by generating the dependent variable Y_i according to (2.6) where

$$h(x) = c_0 \Big[1 - 2\Phi\Big(\frac{x - 1/2}{c_0} \Big) \Big] \quad \text{for some constant} \ \ 0 < c_0 \le 1 \ .$$

The function h is monotonically decreasing where c_0 captures the degree of monotonicity. For small c_0 the function h is close to zero, and for $c_0 = 1$ it holds $h(x) \approx \phi(0)(1 - 2x)$ where ϕ denotes the standard normal probability density function.

We study the size and power patterns of our adaptive ST under the null hypothesis that the NPIV function h is weakly decreasing on the support of X. For the monotonicity test, We implement the test statistic $\widehat{\mathtt{ST}}_n$ given in (2.2) using quadratic B-spline basis functions with varying number of knots for h. Due to piecewise linear derivatives, monotonicity constraints are easily imposed on the restricted function at the derivative at $J - 1$ points. For the basis in the instrument space we also use quadratic B-spline functions with a larger number of knots so that $K(J) = 2J$. We make use of the data-driven critical value $\widehat{\eta}_J(\alpha)$ given in Remark 3.1 (which is implemented using the R package `coneproj`). As a

Sample size	c_0	ξ	*Emp. Size of* $\widehat{\mathtt{ST}}_n$ 10%	5%	1%	*average* $\widehat{J}$ *at 5% level*	*Emp. Size of* $\widehat{\mathtt{ST}}_n^{\mathtt{B}}$ 10%	5%	1%	*average* $\widehat{J}^{\mathtt{B}}$ *at 5% level*
500	0.01	0.3	0.083	0.044	0.009	3.67	0.086	0.046	0.013	3.69
		0.5	0.083	0.044	0.007	3.86	0.083	0.045	0.016	3.86
		0.7	0.082	0.042	0.010	4.05	0.081	0.049	0.013	4.06
	0.1	0.3	0.086	0.045	0.011	3.69	0.088	0.050	0.016	3.71
		0.5	0.085	0.045	0.007	3.91	0.087	0.046	0.018	3.91
		0.7	0.082	0.043	0.011	4.14	0.082	0.047	0.016	4.14
	1	0.3	0.096	0.050	0.013	3.73	0.095	0.055	0.015	3.74
		0.5	0.088	0.045	0.010	3.98	0.088	0.047	0.017	3.97
		0.7	0.084	0.045	0.010	4.25	0.092	0.056	0.015	4.23
1000	0.01	0.3	0.083	0.045	0.009	3.69	0.085	0.041	0.012	3.70
		0.5	0.088	0.047	0.010	3.96	0.086	0.052	0.013	3.98
		0.7	0.084	0.046	0.011	4.45	0.087	0.045	0.015	4.46
	0.1	0.3	0.089	0.049	0.010	3.72	0.087	0.046	0.014	3.74
		0.5	0.091	0.049	0.012	4.03	0.082	0.052	0.014	4.03
		0.7	0.084	0.050	0.014	4.56	0.088	0.049	0.016	4.57
	1	0.3	0.097	0.055	0.011	3.77	0.096	0.053	0.018	3.80
		0.5	0.099	0.056	0.015	4.16	0.088	0.057	0.017	4.14
		0.7	0.104	0.059	0.015	4.53	0.110	0.058	0.019	4.52

Table 1: Testing Monotonicity - Empirical Size for the adaptive tests $\widehat{\mathtt{ST}}_n$ and $\widehat{\mathtt{ST}}_n^{\mathtt{B}}$

comparison, we also implement our adaptive ST using bootstrap critical values as described in Remark 3.2. In each Monte Carlo iteration, we generate 200 bootstrap replications using random weights $\omega \sim \mathcal{N}(0,1)$ drawn independently from (X, W, U).

Table 1 reports the empirical size control for different nominal levels of the test $\widehat{\mathtt{ST}}_n$ and the bootstrap analog $\widehat{\mathtt{ST}}_n^{\mathtt{B}}$ for the sample sizes $n \in \{500, 1000\}$. Results are presented under the different parameter values for $\xi \in \{0.3, 0.5, 0.7\}$ and $c_0 \in \{0.01, 0.1, 1\}$. Overall, we see from Table 1 that the test provides adequate size control for different parameter values of ξ and c_0. It is interesting to see that the computationally demanding bootstrap version $\widehat{\mathtt{ST}}_n^{\mathtt{B}}$ with 200 bootstrap replications does not have any improvement in terms of size control.[5]

In addition to the coverage, Table 1 also presents the average data driven choice of tuning parameter J at the 5% nominal level, denoted by $\widehat{J}$ for the test $\widehat{\mathtt{ST}}_n$ and $\widehat{J}^{\mathtt{B}}$ for its bootstrap analog $\widehat{\mathtt{ST}}_n^{\mathtt{B}}$. Specifically, $\widehat{J}$ is the average choice of J which maximizes $\widehat{\mathcal{W}}_J$ over the index set $\widehat{\mathcal{I}}_n$ when the null is not rejected; and is the smallest $J \in \widehat{\mathcal{I}}_n$ such that $\widehat{\mathcal{W}}_J > 1$ when the null is rejected. (This data-driven choice of J corresponds to *early stopping* when

[5] We did run 1000 Monte Carlo replications with $n = 500$ sample size and 500 bootstrap evaluations per Monte Carlo replications for the monotonicity test. $\widehat{\mathtt{ST}}_n^{\mathtt{B}}$ with 500 bootstrap evaluations have slightly more accurate sizes than that with 200 bootstrap evaluations. Nevertheless, our simple adaptive $\widehat{\mathtt{ST}}_n$ test also has very good size control in 1000 Monte Carlo replications and is super fast to compute.

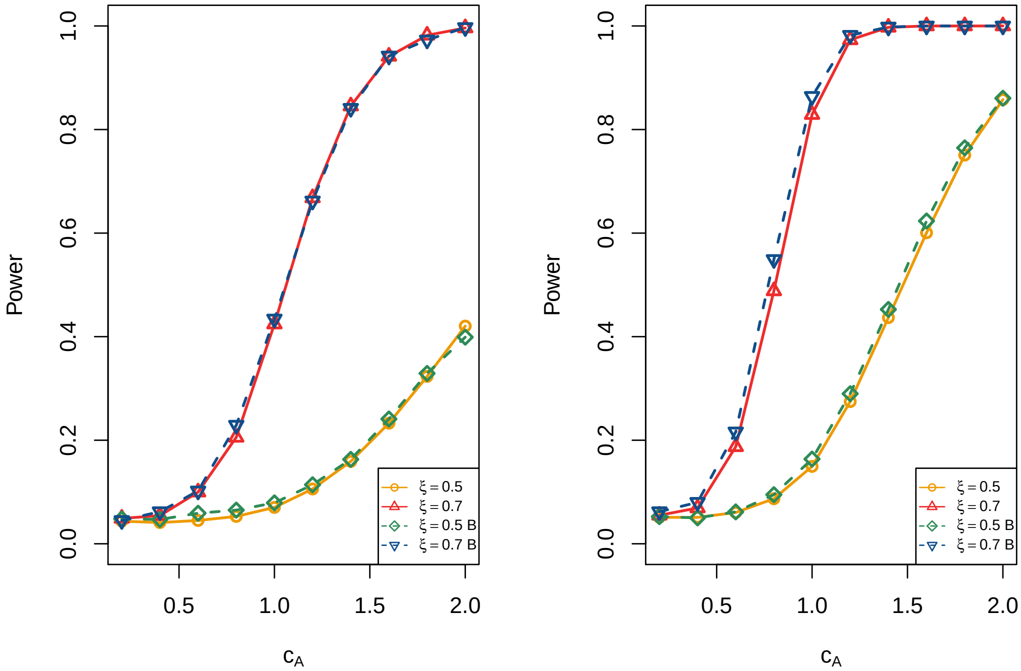


Figure 1: Adaptive Monotonicity Test - Empirical Power for the adaptive tests $\widehat{\mathtt{ST}}_n$ and $\widehat{\mathtt{ST}}_n^{\mathsf{B}}$ when $\xi = \{0.5, 0.7\}$. Solid (or dashed) lines show power results for data-driven (or bootstrap) critical values. Power curves are not size adjusted. LHS: $n = 500$; RHS: $n = 1000$

we reject the null.) From Table 1 we see that the average data driven choice $\widehat{J}$ increases as the strength of instruments increases (captured by the parameter ξ). Further, $\widehat{J}$ decreases as the regularity of the structural function h declines (captured by the parameter c_0). This is due to the fact that with increasing nonlinearity of h a smaller degree of knots is sufficient in order to reject the hypothesis. The data driven choice of J hence works in the opposite direction as in adaptive estimation where larger smoothness leads to smaller values of J. Finally, we see that as the sample size increases so does the value of the estimator $\widehat{J}$.

To study the power of the test that the NPIV function h is monotonically decreasing, we consider deviations from the constant zero function. Specifically, we examine the rejection probabilities of the adaptive ST when the data is generated by the design (2.7) but using the structural function

$$h(x) = -x/5 + c_A x^2.$$

Note that $h'(x) \leq 0$ holds if and only if $x \leq 0.1/c_A$. Since the support of X is contained in $[0, 1]$ we obtain from our model that the null hypothesis of weakly decreasing functions is satisfied only if $c_A \leq 0.1$.

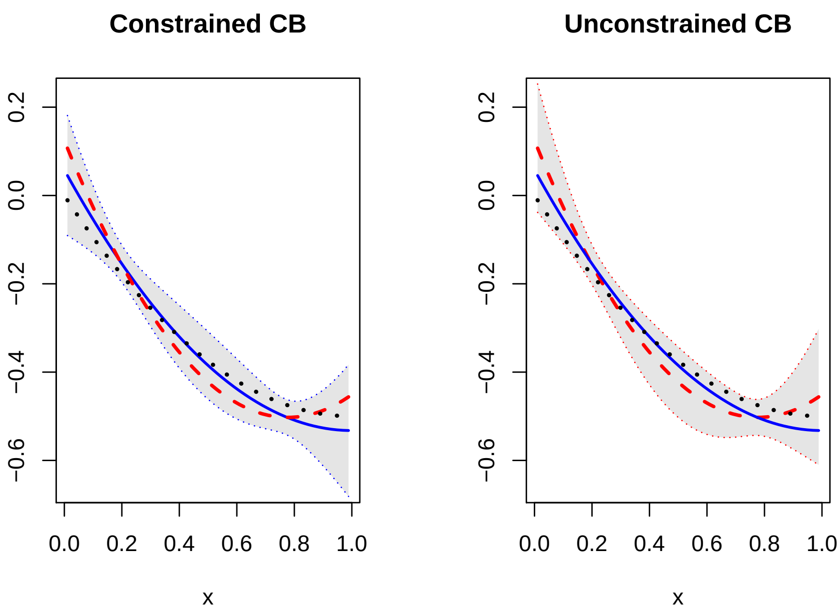


Figure 2: Estimated NPIV curves with data generated from (2.8) with $c_A = 0.1$, $n = 1000$, showing true structural function (black dotted lines), unconstrained estimator (red dashed lines), and constrained estimator (blue solid line). LHS: 95% CB based on constrained estimator. RHS: 95% CB based on unconstrained estimator.

Figure 1 depicts the power function of the adaptive monotonicity test $\widehat{\mathtt{ST}}_n$ and $\widehat{\mathtt{ST}}_n^{\mathtt{B}}$, based on 200 bootstrap iterations, under a 5% nominal level for different parameters $\xi \in \{0.5, 0.7\}$ and sample sizes $n \in \{500, 1000\}$. From Figure 1 we see that both tests become more powerful, for $c_A > 0.1$, as the parameter of instrument strength ξ and the sample size n increase. The bootstrap test $\widehat{\mathtt{ST}}_n^{\mathtt{B}}$ is more powerful for $c_A < 1$ when $\xi = 0.7$ and $n = 1000$. In the other cases, the power improvement by using bootstrap critical values is only of small magnitude or absent. When the power curves are size adjusted the slight advantage in power of $\widehat{\mathtt{ST}}_n^{\mathtt{B}}$ over $\widehat{\mathtt{ST}}_n$ disappears. In particular, we found that $\widehat{\mathtt{ST}}_n$ is more powerful when $c_A > 1$ and $\xi = 0.7$ under size adjustment. Finally, note that we did not report results of $\widehat{\mathtt{ST}}_n^{\mathtt{B}}$ with larger number of bootstrap iterations as it is computationally demanding.

To illustrate the choice of our adaptive inference procedure we further provide implications on estimation. We consider here a modification of our data generating process by considering the model

$$Y = -X + c_A X^2 + U/4 \tag{2.8}$$

where the vector (X, W, U) is generated as in (2.7) and we consider $c_A \in \{0.1, 0.2\}$. For $c_A = 0.1$ the null hypothesis of weakly decreasing structural functions is satisfied and it is violated for $c_A = 0.2$.

We apply our adaptive ST to one sample $\{(Y_i, X_i, W_i)\}$ of size $n = 1000$ for which we obtain a data-driven choice of sieve dimension $\widehat{J} = 3$ in both cases $c_A \in \{0.1, 0.2\}$. Based on the dimension parameter choice, Figure 2 shows the constrained sieve NPIV estimator (blue solid line) and unconstrained sieve NPIV estimator (red dashed lines). We show the 95% uniform confidence bands (CB) following Chen and Christensen [2018] based on 1000 bootstrap on the constrained on the left and based on the unconstrained estimator on the right. From Figure 2 we see that the difference between CBs based on constrained and unconstrained estimator is minor, although there exists a slight improvement of the CB based on the constrained one.

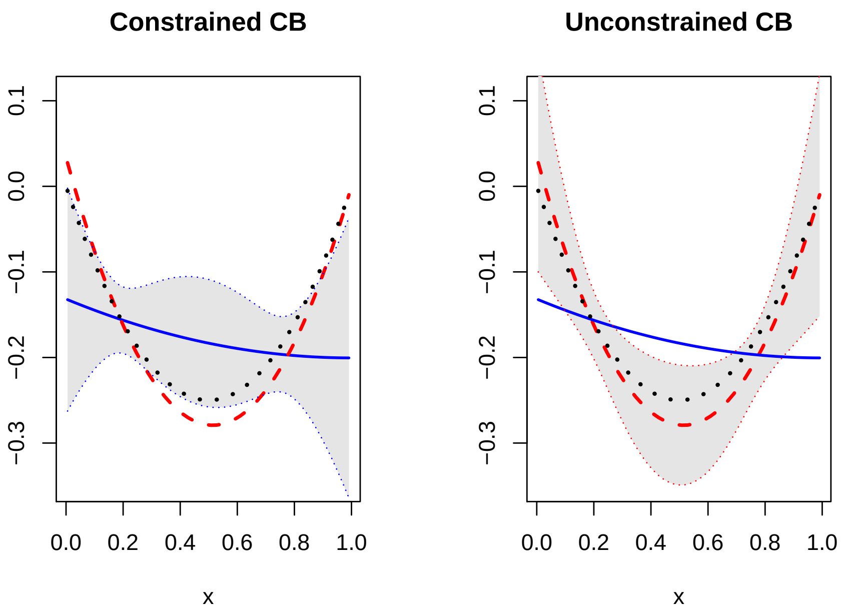


Figure 3: Estimated NPIV curves with data generated from (2.8) with $c_A = 0.2$, $n = 1000$, showing true structural function (black dotted lines), unconstrained estimator (red dashed lines), and constrained estimator (blue solid line). LHS: 95% CB based on constrained estimator. RHS: 95% CB based on unconstrained estimator.

Figure 3 shows the estimation results when $c_A = 0.2$ and hence, the null hypothesis of decreasing curves is violated. While imposing monotonicity constraint has become increasingly popular in NPIV estimation, we see that imposing a wrong shape-constraint can lead to a severe bias.

3. Adaptive Inference via the Structure-Space Test

This section presents several results on data driven structure-space test (ST) statistics. Subsection 3.1 introduces the notation and the main regularity conditions. Subsection 3.2

establishes the minimax rate of testing without a data driven choice of the sieve dimension. Subsection 3.3 establishes the minimax rate of testing of the adaptive ST. Subsection 3.4 shows that this rate coincides with rate of testing attained by the tests with composite null hypothesis. Subsection 3.5 proposes data-driven confidence sets by inverting the adaptive ST under the null hypothesis.

3.1. Main Assumptions

Before we state the minimax rate of testing in structure space, we introduce additional notation and main assumptions. For a random variable X, we define the space $L^2(X)$ as the equivalence class of all measurable functions of X with finite second moment with $\|\cdot\|_{L^2(X)}$ as the associated norm. For any sigma-finite measure μ we define $\|\phi\|_\mu^2 := \int \phi^2(x)\mu(x)dx$ for all $\phi \in L^2_\mu := \{\phi : \|\phi\|_\mu < \infty\}$.

Assumption 1. *(i) $\mathcal{H} \subset L^2(X)$; (ii) $\sup_{w\in\mathcal{W}} \sup_{h\in\mathcal{H}} \mathrm{E}_h[\rho^2(Y, h^{\mathrm{R}}(X))|W = w] \leq \overline{\sigma}^2 < \infty$ and $\sup_{h\in\mathcal{H}} \mathrm{E}_h[\rho^4(Y, h^{\mathrm{R}}(X))] < \infty$; and (iii) $\inf_{w\in\mathcal{W}} \inf_{h\in\mathcal{H}} \mathbb{V}ar_h(\rho(Y, h^{\mathrm{R}}(X))|W = w) \geq \underline{\sigma}^2 > 0$.*

Let $A = [S'G_b^{-1}S]^{-1}S'G_b^{-1}$ where $S = \mathrm{E}[b^K(W)\psi^J(X)']$ and $G_b = \mathrm{E}[b^K(W)b^K(W)']$. We introduce the projections $\Pi_J h(\cdot) = \psi^J(\cdot)'G^{-1}\int \psi^J(x)h(x)\mu(x)dx$ for $h \in L^2_\mu$ and $\Pi_K m(\cdot) = b^K(\cdot)'G_b^{-1}\,\mathrm{E}[b^K(W)m(W)]$ for $m \in L^2(W)$. The minimal singular value of $G_b^{-1/2}SG^{-1/2}$ is denoted by s_J. We make use of the notation $\zeta_J = \max(\zeta_{\psi,J}, \zeta_{b,K})$, for $K = K(J)$, where $\zeta_{\psi,J} = \sup_x \|G^{-1/2}\psi^J(x)\|$ and $\zeta_{b,K} = \sup_w \|G_b^{-1/2}b^K(w)\|$. We define $d_x = \dim(X)$ and $d_w = \dim(W)$.

Assumption 2. *(i) $s_J^{-1}\zeta_J^2\sqrt{(\log J)/n} = O(1)$; (ii) $\|\Pi_J h - h\|_\mu = O(J^{-p/d_x})$ for all $h \in \mathcal{H}$ and some $p > 0$ such that $\zeta_J\sqrt{\log J} = O(J^{p/d_x})$; and (iii) the eigenvalues of G and G_b are uniformly bounded from below and above.*

Let $T : L^2(X) \mapsto L^2(W)$ denote the conditional expectation operator given by $Th(w) = \mathrm{E}[h(X)|W = w]$. We further define $\Psi_J = \mathrm{clsp}\{\psi_1, \ldots, \psi_J\} \subset L^2(X)$.

Assumption 3. *(i) $\sup_{h\in\Psi_J} \|(\Pi_K T - T)h\|_{L^2(W)}/\|h\|_\mu = o(s_J)$ and (ii) $\|T(h - h^{\mathrm{R}} - \Pi_J(h - h^{\mathrm{R}})\|_{L^2(W)} = O(s_J\|h - h^{\mathrm{R}} - \Pi_J(h - h^{\mathrm{R}})\|_\mu)$ for all $h \in \mathcal{H}$.*

Assumption 4. *For any $h \in \mathcal{H}$, $Th = 0$ implies that $\|h\|_\mu = 0$.*

Discussion of Assumptions. Assumption 1 captures second moment bounds. In addition, a lower bound for the variance is imposed. Assumption 2 (i) imposes bounds on the growth of the basis functions relative to the singular values of the matrix $G_b^{-1/2}SG^{-1/2}$. Assumption 2 (i)(ii) imposes bounds on the growth of the basis functions which are known for commonly used bases. For instance, $\zeta_{b,K} = O(\sqrt{K})$ and $\zeta_{\psi,J} = O(\sqrt{J})$ for polynomial spline, wavelet and cosine bases, and $\zeta_{b,K} = O(K)$ and $\zeta_{\psi,J} = O(J)$ for orthogonal polynomial bases; see, e.g., Newey [1997], Huang [1998]. Assumption 3 (i) is a mild condition on the approximation properties of the basis used for the instrument space. In fact, $\|(\Pi_K T - T)h\|_{L^2(W)} = 0$ for all $h \in \Psi_J$ when the basis functions for B_K and Ψ_J form either a Riesz basis or an eigenfunction basis for the conditional expectation operator. Assumption 3 (ii) is the usual L^2 "stability condition" imposed in the NPIV literature when $h^{\mathrm{R}} = 0$ (cf. Assumption 6 in Blundell et al. [2007] and Assumption 5.2(ii) in Chen and Pouzo [2012]). Note that Assumption 3 (ii) is also automatically satisfied by Riesz bases. Assumption 4 is required for identification of the quadratic functional $\|h\|_\mu$ and the condition can be less restrictive than imposing L^2 completeness when the support of μ is a subset of the support of X.

Example 3.1 (NQIV)**.** *The ST test can also be applied to models with nonseperable unobservables after linearization. Consider as an example the nonparametric quantile instrumental variable model with conditional moment restriction*

$$\mathrm{E}[\mathbb{1}\{Y \le h(X)\} - q|W] = 0$$

for some $q \in (0,1)$. A linearization of the model can be obtained using the Frechet derivative at h^{R} maps h to $\mathrm{E}[f_{Y|X,W}(h^{\mathrm{R}}(X))(h(X) - h^{\mathrm{R}}(X))|W]$. This leads to a modified version of our test statistic where $B'\Psi$ is replaced by an empirical analog of $\mathrm{E}[f_{Y|X,W}(h^{\mathrm{R}}(X))b^K(W)\psi^J(X)']$ and $Y - h^{\mathrm{R}}(X)$ by $f_{Y|X,W}(h^{\mathrm{R}}(X))h^{\mathrm{R}}(X)$. We do not address the estimation of the conditional density and hence, the NQIV case explicitly for our structural space test.

Example 3.2 (Testing Derivatives of h)**.** *Note that the test can be extended to check for derivatives of the function h. To do so, we replace G by the matrix*

$$\int \partial_x \psi^J(x)(\partial_x \psi^J)(x)'\mu(x)dx$$

as long as the basis function ψ_j are differentiable on the support of μ. This straightforward extension is only possible in case of ST but not for IT and hence, illustrates the advantage of the ST approach.

3.2. The Minimax Rate of ST Under Simple Null Hypothesis

We first consider the simple hypothesis case where $\mathcal{H}_0 = \{h_0\}$ and, in particular, $h^{\mathrm{R}} = h_0$, for some known function h_0 satisfying (1.1) with $\rho(Y, h(X)) = Y - h(X)$. We introduce a J dependent analog to the adaptive structure-space test $\widehat{\mathtt{ST}}_n$ under the simple null:

$$\mathtt{ST}_{n,J} = \mathbb{1}\left\{ \frac{n\widehat{D}_J(h_0)}{\sqrt{2}\,\eta\,\widehat{\mathrm{v}}_J(h_0)} > 1 \right\}$$

for some constant $\eta > 0$. The test $\mathtt{ST}_{n,J}$ with optimally chosen J serves as a benchmark of our adaptive ST procedure (given in (3.5)) for the simple null hypothesis case.

3.2.1. Upper bound

Theorem 3.1. *Let Assumptions 1–4 be satisfied. Then, for any $\varepsilon > 0$ there exists a constant $\delta^* > 0$ such that*

$$\limsup_{n\to\infty} \Big\{ \mathrm{P}_{h_0}(\mathtt{ST}_{n,J} = 1) + \sup_{h\in\mathcal{H}_1(\delta^*, r_{n,J})} \mathrm{P}_h(\mathtt{ST}_{n,J} = 0) \Big\} \leq \varepsilon, \tag{3.1}$$

where the rate $r_{n,J}$ is given by

$$r_{n,J} = n^{-1/2} s_J^{-1} J^{1/4} + J^{-p/d_x}. \tag{3.2}$$

Theorem 3.1 shows that the test statistic $\mathtt{ST}_{n,J}$ attains the rate of testing $r_{n,J}$. This rate consists of a variance and a bias part. The optimal choice of J requires knowledge of unknown mapping properties of the conditional expectation operator T and the unknown smoothness of the true structural function h, as illustrated below. A central step to achieve this rate result is to establish a rate of convergence of the quadratic distance estimator $\widehat{D}_J(h_0)$, see Theorem F.1 in the online appendix. We thus make use of the close connection between minimax optimal quadratic functional estimation and minimax optimal testing.

We differentiate among two different degrees of ill-posedness, which are typically considered in the literature. The sieve L^2 measure of ill-posedness is defined as

$$\tau_J = \sup_{h\in\Psi_J, h\neq 0} \frac{\|h\|_\mu}{\|Th\|_{L^2(W)}} \leq \sup_{h\in\Psi_J, h\neq 0} \frac{\|h\|_\mu}{\|\Pi_K Th\|_{L^2(W)}} = s_J^{-1}.$$

We call the model (1.1) mildly ill-posed if: $\tau_j \sim j^{\zeta/d_x}$ for some $\zeta > 0$ and severely ill-posed if: $\tau_j \sim \exp(j^{\zeta/d_x}/2)$ for some $\zeta > 0$.[6] The next corollary provides concrete rates of testing when the dimension parameter J is chosen to level variance and square bias under classical smoothness conditions.

Corollary 3.1. *Let Assumptions 1–4 be satisfied. Then the rate of testing $r_{n,J}$ given in (3.2) is of the following form:*

1. *Mildly ill-posed case: choosing $J \sim n^{2d_x/(4(p+\zeta)+d_x)}$ implies*

$$r_{n,J} = n^{-2p/(4(p+\zeta)+d_x)}, \tag{3.3}$$

[6] If $\{a_n\}$ and $\{b_n\}$ are sequences of positive numbers, we use the notation $a_n \lesssim b_n$ if $\limsup_{n\to\infty} a_n/b_n < \infty$ and $a_n \sim b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

2. *Severely ill-posed case: choosing* $J \sim \big(\log n - \frac{2p+d_x}{\zeta}\log\log n\big)^{d_x/\zeta}$ *implies*

$$r_{n,J} = (\log n)^{-p/\zeta}. \tag{3.4}$$

3.2.2. Lower bound

In the next result, we establish a lower bound for the rate of testing in each of the ill-posed case scenarios considered in the previous corollary. Below, $\langle \cdot, \cdot \rangle_\mu$ denotes the inner product associated to L^2_μ.

Theorem 3.2. *Let Assumptions 1 (iii) and 4 be satisfied. Assume that* $\|Th\|^2_{L^2(W)} \lesssim \sum_{j\geq 1} \tau_j^{-2} \langle h, \widetilde{\psi}_j \rangle^2_\mu$ *for all* $h \in \mathcal{H}$ *and an orthonormal basis* $\{\widetilde{\psi}_j\}_{j\geq 1}$ *in* L^2_μ*. Then for any* $\varepsilon > 0$ *there exists a constant* $\delta_* > 0$ *such that*

$$\liminf_{n\to\infty} \inf_{\mathtt{T}_n} \Big\{ \mathrm{P}_{h_0}(\mathtt{T}_n = 1) + \sup_{h \in \mathcal{H}_1(\delta_*, r_n)} \mathrm{P}_h(\mathtt{T}_n = 0) \Big\} \geq 1 - \varepsilon$$

where r_n *is given by:*

1. *Mildly ill-posed case:* $r_n = n^{-2p/(4(p+\zeta)+d_x)}$*,*
2. *Severely ill-posed case:* $r_n = (\log n)^{-p/\zeta}$*.*

From Corollary 3.1 and Theorem 3.2 we conclude that $r_{n,J}$ is the *minimax rate of testing* once J is chosen to level variance and squared bias. In particular, we conclude that the rate of testing is always nonparametric in contrast to the case of estimation of quadratic functionals where also the $\sqrt{n}$–rate can be achieved.

3.3. Adaptive ST Under Simple Null Hypothesis Case

We propose a data-driven ST that rejects the null hypothesis $\mathcal{H}_0 = \{h_0\} \neq \emptyset$, for some known function h_0 satisfying (1.1), as soon as at least for one J the normalized estimator $\widehat{D}_J(h_0)$ is sufficiently large. Specifically, we consider the data-driven test statistic

$$\mathtt{ST}_n = \mathbb{1}\left\{ \exists J \in \widehat{\mathcal{I}}_n \text{ such that } \frac{n\widehat{D}_J(h_0)}{\sqrt{2}\widehat{\eta}_J(\alpha)\widehat{\mathrm{v}}_J(h_0)} > 1 \right\}, \tag{3.5}$$

where $\widehat{\eta}_J(\alpha)$, $\widehat{\mathrm{v}}_J(h_0)$, and the index set $\widehat{\mathcal{I}}_n$ are given in Subsection 2.2.

We define J_0 be the smallest dimension parameter such that the variance dominates the squared bias within a $\sqrt{\log\log n}$ term, that is,

$$J_0 = \min\big\{ J : \ J^{-2p/d_x} \leq n^{-1}\sqrt{\log\log n}\, s_J^{-2}\sqrt{J} \big\}. \tag{3.6}$$

Recall the definition of the index set $\widehat{\mathcal{I}}_n$ given in (2.4) which relies on an upper bound $\widehat{J}_{\max}$. We introduce the dimension parameter $\overline{J}$ slowly growing with the sample size n which controls the complexity of the ES index set $\widehat{\mathcal{I}}_n$. Specifically, Chen and Christensen [2015, Theorem 3.2] show that $\widehat{J}_{\max} \leq \overline{J}$ holds with probability approaching one where $\overline{J}$ satisfies the rate restrictions imposed in the next assumption. We also make use of the notation $\overline{\zeta} = \zeta_{\overline{J}}$.

Assumption 5. *(i)* $s_{\overline{J}}^{-1}\overline{\zeta}^2\sqrt{(\log n)/n} = o(1)$; *(ii) for any* $\alpha \in (0,1)$ *it holds* $\widehat{\eta}_{\overline{J}}(\alpha) = O(\sqrt{\log\log n})$ *and* $(\log\log J)^c \leq \widehat{\eta}_J(\alpha)$ *for some constant* $c > 1$ *and for all* $\underline{J} \leq J \leq \overline{J}$ *with probability approaching one.*

Assumption 5 imposes an upper bound on the growth of the population counterpart of the upper bound of the set $\widehat{\mathcal{I}}_n$. Assumption 5 (i) is a slight modification of Assumption 2 (i) considered uniformly over $\underline{J} \leq J \leq \overline{J}$. Assumption 5 (ii) imposes a mild restriction on the critical values $\widehat{\eta}_J(\alpha)$ given in (2.5).

Theorem 3.3. *Let Assumptions 1–3 and 5 be satisfied. Then, for any* $\varepsilon > 0$ *there exists a constant* $\delta^\circ > 0$ *such that*

$$\limsup_{n\to\infty}\Big\{\mathrm{P}_{h_0}(\mathtt{ST}_n = 1) + \sup_{h\in\mathcal{H}_1(\delta^\circ, \mathsf{r}_n)} \mathrm{P}_h(\mathtt{ST}_n = 0)\Big\} \leq \varepsilon, \tag{3.7}$$

where the rate r_n *is given by*

$$\mathsf{r}_n^2 = n^{-1}\sqrt{\log\log n}\, s_{J_0}^{-2}\sqrt{J_0}. \tag{3.8}$$

Theorem 3.3 establishes an upper bound for the testing rate of the adaptive structure-space test $\mathtt{ST}_n$. The proof of Theorem 3.3 relies on a novel exponential bound for degenerate U-statistics based on sieve estimators. Adaptive testing for inverse problems was considered for deconvolution models (with known degree of ill-posedness) by Butucea et al. [2009]. In functional linear models, adaptive tests (under unknown, mild degree of ill-posedness) were proposed by Lei [2014]. In Gaussian white noise models, adaptive tests was proposed by Ingster et al. [2012] also under the severely ill-posed case but requires knowledge of the ill-posedness scenario. We now illustrate the upper bound under classical smoothness assumptions. Again, we distinguish between the mildly or severely ill-posed case.

Corollary 3.2. *Let Assumptions 1–5 be satisfied. Then, the adaptive rate of testing* r_n *given in* (3.8) *satisfies:*

1. *Mildly ill-posed case:*

$$\mathsf{r}_n = \big(\sqrt{\log\log n}/n\big)^{2p/(4(p+\zeta)+d_x)},$$

2. *Severely ill-posed case:*

$$r_n = (\log n)^{-p/\zeta}.$$

From Corollary 3.2 we see that the adaptive ST attains in the mildly ill-posed case the minimax rate of testing within a $(\log\log n)$-term. For adaptive testing without endogeneity, it is well known that a $(\log\log n)$-term is required, see Spokoiny [1996]. In the severely ill-posed cases, our adaptive test attains the exact minimax rate of testing and hence, there is no price to pay for adaptation.

3.4. Adaptive Testing Under Composite Null Hypothesis Case

We extend the results from the previous subsection to the case of composite hypotheses and, in particular, allow for testing inequality constraints. Below, we discuss two different approaches for deriving critical values in the case of constrained inequality tests. Both methods rely on cone properties imposed on the restricted set of functions. We use the notation Π_J^{R} for the projection on $\mathcal{H}_J^{\mathrm{R}}$. Here, the set $\mathcal{H}_J^{\mathrm{R}}$ is used to approximate the set of functions $\mathcal{H}^{\mathrm{R}} \subset \mathcal{H}$ which satisfies a conjectured restriction.

Remark 3.1 (Adaptive critical values for inequality constrains)**.** *In both cases, we rely on the assumption that $\mathcal{H}_J^{\mathrm{R}}$ is a polyhedral cone*[7]*. In this case, we may infer from Silvapulle and Sen [2005, Lemma 3.13.5] the existence of a collection of faces $\{H_1, \dots, H_L\}$ such that the collection of their relative interiors $\{ri(H_1), \dots, ri(H_L)\}$ forms a partitioning of $\mathcal{H}_J^{\mathrm{R}}$. Let P_l be the projection matrix onto the linear space spanned by H_l where $J_l = rank(P_l)$. Then, a Bonferroni correction of the adaptive critical values of Al Mohamad et al. [2018] gives*

$$\widehat{\eta}_J(\alpha) = \sum_{l=1}^{L} \mathbb{1}\big\{\Pi_J^{\mathrm{R}} \widehat{h}_J \in ri(H_l)\big\} \frac{q\big(\alpha/\#(\widehat{\mathcal{I}}_n), J_l\big) - J_l}{\sqrt{2J_l}},$$

where $\widehat{h}_J$ is the unconstrained analog of (2.1) and we impose the restriction $J_l \geq 1$.

Remark 3.2 (Bootstrap critical values for inequality constrains)**.** *We propose a modification of the bootstrap procedure of Fang and Seo [2019] which imposes a cone condition on $\mathcal{H}^{\mathrm{R}}$. Below $\mathbb{Z}_J$ denotes the sieve bootstrap score proposed by Chen and Christensen [2018]. We proceed in two steps:*
STEP 1. *Introduce a sequence of independent and identically distributed random variables $\{\omega_i\}_{i=1}^n$ drawn independently of the original data $\{(Y_i, X_i, W_i)\}_{i=1}^n$. Compute the bootstrap*

[7] A cone $\mathcal{C}$ is called polyhedral if there is some matrix M such that $\mathcal{C} = \{\beta \in \mathbb{R}^J : M\beta \geq 0\}$.

version of the quadratic distance estimator given by

$$\widehat{D}^{\mathtt{B}}_J(\widehat{h}_J) = \frac{2}{n(n-1)} \sum_{1\leq i<i'\leq n} \widehat{U}_i\widehat{U}_{i'} b^{K(J)}(W_i)'\widehat{A}'G\widehat{A}\, b^{K(J)}(W_{i'}), \tag{3.9}$$

where $\widehat{U}_i = \omega_i(Y_i - \widehat{h}_J(X_i))$.

STEP 2. *For some* $\gamma_n \in (0,1)$,[8] *we construct the* $1-\gamma_n$ *quantile of* $\{n\widehat{D}^{\mathtt{B}}_J(\widehat{h}_J)/\sqrt{2}\widehat{\mathrm{v}}_J(\widehat{h}_J)\}$, *based on* N *bootstrap samples, denoted by* $\widehat{\tau}_{n,1-\gamma_n}$. *Set* $\widehat{\kappa}_J = \nu_n c_n/\widehat{\tau}_{n,1-\gamma_n}$ *where* $\nu_n^2 = n/\sqrt{2}\widehat{\mathrm{v}}_J(\widehat{h}_J)$ *and* c_n *is such that* $\|(\widehat{h}_J - h)/\sqrt{2}\widehat{\mathrm{v}}_J(\widehat{h}_J) - \mathbb{Z}_J\|_\mu = o_p(c_n)$. *Compute* $\widehat{\eta}_J(\alpha)$ *as the* $1-\alpha/\#(\widehat{\mathcal{I}}_n)$ *quantile of* $\{n\widetilde{D}^{\mathtt{B}}_J(\widehat{h}_J)/\sqrt{2}\widehat{\mathrm{v}}_J(\widehat{h}_J)\}$ *based on* N *bootstrap samples, where* $\widetilde{D}^{\mathtt{B}}_J(\widehat{h}_J)$ *coincides with* (3.9) *but with* $\widehat{U}_i$ *replaced by* $\widehat{U}_i - \widehat{\kappa}_J\widehat{h}_J(X_i) - \Pi^{\mathrm{R}}_J(\mathbb{Z}_J - \widehat{\kappa}_J\widehat{h}_J)$.

In the following, we impose restrictions on the complexity of the set $\mathcal{H}^{\mathrm{R}}_J$. As we rely on the empirical process theory, we make use of the literature's notation. Let $N_{[]}(t,\mathcal{H},\|\cdot\|_\mu)$ denote the smallest number of brackets of size t (under $\|\cdot\|_\mu$) required to cover $\mathcal{H}$. We further denote $\mathcal{H}^{\mathrm{R}}_J(\Delta_{J,n}) = \{h \in \mathcal{H}^{\mathrm{R}}_J : \|h - h^{\mathrm{R}}\|_\infty \leq \Delta_{J,n}\}$ for some $\Delta_{J,n} > 0$. Below, $\eta_J(\alpha)$ denotes a deterministic sequence satisfying $\eta_{\overline{J}}(\alpha) = O(\sqrt{\log\log n})$ and $(\log\log J)^c \leq \eta_J(\alpha)$ for some constant $c > 1$ and for all $\underline{J} \leq J \leq \overline{J}$.

Assumption 6. *(i) For any* $h \in \mathcal{H}$ *there exists a sequence* $(\Delta_{J,n})_{n\geq 1}$ *satisfying* $\widehat{h}^{\mathrm{R}}_J \in \mathcal{H}^{\mathrm{R}}_J(\Delta_{J,n})$ *with probability approaching one and* $\int_0^1 \sqrt{1 + \log N_{[]}(tC, \mathcal{H}^{\mathrm{R}}_J(\Delta_{J,n}), \|\cdot\|_\mu)}dt \leq C_{J,n}$ *where* $\sum_{J\in\widehat{\mathcal{I}}_n} C^2_{J,n}\Delta^2_{J,n}/(\log\log J) = o_p(1)$ *and* $\max_{J\in\widehat{\mathcal{I}}_n} \Delta^2_{J,n}\zeta^2_J(\log J) = o_p(1)$. *(ii) For some* $J \in \widehat{\mathcal{I}}_n$ *and any* $\varepsilon > 0$ *there exist constants* $c, C > 0$ *such that it holds* $\limsup_{n\to\infty}\sup_{h\in\mathcal{H}_0} \mathrm{P}_h\big(\widehat{\eta}_J(\alpha) < C\,\eta_J(\alpha)\big) < \varepsilon$ *and* $\limsup_{n\to\infty}\sup_{h\in\mathcal{H}_1(\delta^\circ, r_n)} \mathrm{P}_h\big(\widehat{\eta}_J(\alpha) > c\,\eta_J(\alpha)\big) < \varepsilon$.

Assumption 6 (i) is a mild restriction on the complexity of the set of functions $\mathcal{H}^{\mathrm{R}}_J(\Delta_{J,n})$ by imposing rate conditions on $\Delta_{J,n}$. These conditions determine the rate of convergence of the constraint sieve estimator to any function h^{R} satisfying a conjectured restriction captured by $\mathcal{H}^{\mathrm{R}}$. It was similarly imposed in Assumption C.2 by Chen and Pouzo [2012]. Further note that the critical values estimators introduced in Remarks 3.1 and 3.2 satisfy Assumption 6 (ii) under mild conditions. In the case of the adaptive critical values (see Remark 3.1), the cone projection leads to a weakly larger critical values than the one given in (2.5) since $\widehat{\eta}_J(\alpha)$ is now determined by the dimension of the face on which the cone projection lands. In the case of the bootstrap critical values (see Remark 3.2), Assumption 6 (ii) can be justified by following Fang and Seo [2019].

The next result establishes an upper bound for the rate of testing of $\widehat{\mathtt{ST}}_n$.

[8] In the implementation of the procedure we use throughout the paper the choice $\gamma_n = 0.1/\log n$, following Chernozhukov et al. [2013].

Theorem 3.4. *Let Assumptions 1–6 be satisfied. Then, for any $\varepsilon > 0$ there exists a constant $\delta^\circ > 0$ such that*

$$\limsup_{n\to\infty}\Big\{\sup_{h\in\mathcal{H}_0} \mathrm{P}_h\big(\widehat{\mathtt{ST}}_n = 1\big) + \sup_{h\in\mathcal{H}_1(\delta^\circ,\mathsf{r}_n)} \mathrm{P}_h\big(\widehat{\mathtt{ST}}_n = 0\big)\Big\} \leq \varepsilon, \tag{3.10}$$

where the rate r_n is given in Theorem 3.3.

From Theorem 3.4 we see that $\widehat{\mathtt{ST}}_n$ attains the rate of testing r_n which is the same rate of testing obtained by $\mathtt{ST}_n$ in the case of simple hypotheses. Under the restriction imposed in Assumption 6 we thus conclude that estimation of restricted functions does not imply slower rates of testing. In the definition of $\widehat{\mathtt{ST}}_n$, the dimension parameter for estimating the structural function under the conjectured restriction is set to be equivalent to the unrestricted estimator of the structural function. In this sense, our inference results do not require undersmoothing conditions. Finally, we note that the test statistic can be trivially modified to tests where the constraint functions h_0 might be estimated using a fixed finite dimensional sieve space.

3.5. Confidence Sets

We now propose L^2 confidence sets which are based on inversion of the structural space test. The resulting confidence region imposes conjectured restrictions on the function of interest h. The $(1-\alpha)-$confidence set is given by

$$\mathcal{C}_n(\alpha) = \Big\{h \in \mathcal{H}^{\mathrm{R}} : \frac{n\widehat{D}_J(h)}{\sqrt{2}\,\widehat{\eta}_J(\alpha)\widehat{\mathrm{v}}_J(h)} \leq 1 \quad \text{for all } J \in \widehat{\mathcal{I}}_n\Big\}. \tag{3.11}$$

The following corollary exploits our previous results and the introduced assumptions to characterize the asymptotic size and power properties of our procedure.

Corollary 3.3. *Let Assumptions 1–6 be satisfied. Then, for any $\alpha > 0$ it holds*

$$\limsup_{n\to\infty}\sup_{h\in\mathcal{H}_0} \mathrm{P}_h\left(h \notin \mathcal{C}_n(\alpha)\right) \leq \alpha \tag{3.12}$$

and there exists a constant $\delta^\circ > 0$ such that

$$\liminf_{n\to\infty}\inf_{h\in\mathcal{H}_1(\delta^\circ,\mathsf{r}_n)} \mathrm{P}_h\left(h \notin \mathcal{C}_n(\alpha)\right) \geq 1-\alpha. \tag{3.13}$$

Corollary 3.12 shows that the L^2 confidence set $\mathcal{C}_n(\alpha)$ controls size uniformly over the class of functions $\mathcal{H}_0$. Moreover, the result establishes power uniformly over the class $\mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$. We immediately see from Corollary 3.3 that the size of the L^2 confidence ball depends on the degree of ill-posedness captured by the minimal singular values s_J.

Corollary 3.4. *Let Assumptions 1–6 be satisfied. Then, we have*

$$\limsup_{n\to\infty}\sup_{h\in\mathcal{H}} \mathrm{P}_h\left(\textit{diam}(\mathcal{C}_n(\alpha)) \geq Cn^{-1}\sqrt{\log\log(n)}s_{J_0}^{-2}\sqrt{J_0}\right) = 0,$$

for some constant $C > 0$ and where the dimension parameter J_0 is given in (3.6).

Corollary 3.4 yields a confidence region of the diameter of the order $n^{-1}\sqrt{\log\log(n)}s_{J_0}^{-2}\sqrt{J_0}$ for confidence sets based on inversion of the structural space test statistic. We see that the diameter of the confidence set does not adapt to regularity of submodels. The following remark illustrates that the gain of adaptation is expected to minor in inverse problems.

Remark 3.3 (Adaptive Confidence Sets)**.** *Consider the mildly ill-posed case where the degree of ill-posedness is fixed to ζ and we wish to adapt over the function classes $\mathcal{H}(p)$ and $\mathcal{H}(p_1)$ with smoothness parameters $p_1 > p$. Suppose we have an adaptive estimator which attains in the mildly ill-posed case the L^2 rate of convergence $n^{-p/(2(p+\zeta)+d_x)}$ in comparison to the rate of testing $n^{-p/(2(p+\zeta)+d_x/2)}$. It is known in statistical regression models (see Robins and Van Der Vaart [2006] and Cai and Low [2006]) that rate adaption is only possible over submodels (indexed by p_1) such that the rate of estimation over the submodel is larger than the rate of testing over the "supermodel". Specifically, we obtain the restriction*

$$n^{-p/(2(p+\zeta)+d_x/2)} \lesssim n^{-p_1/(2(p_1+\zeta)+d_x)}.$$

This condition translates into the smoothness restriction

$$p_1 < p\frac{2\zeta + d_x}{2\zeta + d_x/2}$$

and hence, in this sense, adaptation is only possible when the submodel $\mathcal{H}(p_1)$ satisfies

$$p_1 \in \left(p, p\frac{2\zeta + d_x}{2\zeta + d_x/2}\right),$$

which shows that for large values of ζ (or dimension of d_x) adaptation with respect to $\mathcal{H}(p_1)$ can only be achieved over a very limited range of smoothness p_1.

4. Adaptive Inference via the Image-Space Test

In this section we consider a data-driven test in the image-space of the conditional expectation mapping. Subsection 4.1 proposes a data-driven image-space test (IT) statistic. Subsection 4.2 establishes the minimax rate of testing of the adaptive IT.

4.1. The Data-driven Image-Space Test Statistic

We consider the set of function satisfying $\mathcal{H}_0 = \{h \in \mathcal{H} : m(\cdot, h) = 0\} \cap \mathcal{H}^{\mathrm{R}} \neq \emptyset$ where, in this section, $\mathcal{H}^{\mathrm{R}}$ is a finite dimensional, compact function space. We do not address the question of IT with infinite dimensional restricted set of functions here as this would require an additional choice of tuning parameter.

In contrast to the previous section, we specify alternative models through deviations from the conditional moment restriction. For convenience of notation, we introduce the conditional moment function $m(\cdot, h^{\mathrm{R}}) = \mathrm{E}[\rho(Y, h^{\mathrm{R}}(X))|W = \cdot]$. For the image space test, we consider a class of functions which are separated from $\mathcal{H}_0$ in the sense

$$\mathcal{M}_1(\delta, r_n) := \Big\{ h \in \mathcal{H} : m(W, h) = 0 \quad \text{and} \quad \mathrm{E}[m^2(W, h^{\mathrm{R}})] \geq \delta\, r_n^2 \Big\}.$$

We propose an image-space test based on a leave-one-out sieve estimator of the quadratic functional $\mathrm{E}[m^2(W, h^{\mathrm{R}})]$ given by

$$\widehat{D}_K(\widehat{h}^{\mathrm{R}}) = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \rho\big(Y_i, \widehat{h}^{\mathrm{R}}(X_i)\big)\rho\big(Y_{i'}, \widehat{h}^{\mathrm{R}}(X_{i'})\big) b^K(W_i)'(B'B/n)^{-}\, b^K(W_{i'}).$$

where $\widehat{h}^{\mathrm{R}}$ is a constrained estimator given by

$$\widehat{h}^{\mathrm{R}} = \operatorname*{arg\,min}_{h \in \mathcal{H}^{\mathrm{R}}} \sum_{1 \leq i, i' \leq n} \rho\big(Y_i, h(X_i)\big)\rho\big(Y_{i'}, h(X_{i'})\big) b^K(W_i)'(B'B/n)^{-}\, b^K(W_{i'}). \tag{4.1}$$

We reject the null hypothesis as soon as at least for one K the normalized estimator $\widehat{D}_K(\widehat{h}^{\mathrm{R}})$ is sufficiently large. The data-driven image-space test (IT) statistic is given by

$$\widehat{\mathtt{IT}}_n = \mathbb{1}\left\{ \exists K \in \widehat{\mathcal{I}}_n \text{ such that } \frac{n\widehat{D}_K(\widehat{h}^{\mathrm{R}})}{\sqrt{2}\eta_K(\alpha)\widehat{\mathrm{v}}_K(\widehat{h}^{\mathrm{R}})} > 1 \right\}, \tag{4.2}$$

based on $\eta_K(\alpha)$ as given in (2.5) and the estimated normalization factor

$$\widehat{\mathrm{v}}_K(h) = \Big\| n^{-1} \sum_{i=1}^{n} \rho^2\big(Y_i, h(X_i)\big)(B'B/n)^{-1/2}\, b^K(W_i) b^K(W_i)'(B'B/n)^{-1/2} \Big\|_F. \tag{4.3}$$

Also the image space test $\mathtt{IT}_n$ relies on the ES selection method to determine the index set $\widehat{\mathcal{I}}_n$ as given in (2.4) yet its upper bound has to be modified as follows. We replace the upper bound $\widehat{J}_{\max}$ of the index set $\widehat{\mathcal{I}}_n$ by the estimator

$$\widehat{K}_{\max} = \min\Big\{ K > \underline{J} : \zeta^2(K)\sqrt{\ell(K)(\log n)/n} \geq s_{\min}\big((B'B/n)^{-1/2}\big) \Big\}$$

where $\ell(K) = 0.1 \log\log K$.

4.2. Adaptive Testing

For the IT case the index set $\widehat{\mathcal{I}}_n$ depends on the upper bound $\widehat{K}_{\max}$. As in the previous section, we may assume that $\widehat{K}_{\max} \leq \overline{K}$ with probability approaching one where $\overline{K}$ satisfies the following rate conditions. We further denote $\mathcal{H}^{\mathrm{R}}(\Delta_n) = \{h \in \mathcal{H}^{\mathrm{R}} : \|h - h^{\mathrm{R}}\|_\infty \leq \Delta_n\}$ for some $\Delta_n > 0$, where $h^{\mathrm{R}} = \arg\min_{h \in \mathcal{H}^{\mathrm{R}}} \|m(\cdot, h)\|_{L^2(W)}$. Further, we define $\widetilde{b}^K = G_b^{-1/2} b^K$ with entries $\widetilde{b}_k$, $1 \leq k \leq K$.

Assumption 7. *(i)* $\zeta_{b,\overline{K}}^2 \sqrt{(\log n)/n} = o(1)$ *(ii)* $\|\Pi_K m(\cdot, h^{\mathrm{R}}) - m(\cdot, h^{\mathrm{R}})\|_{L^2(W)} = O(K^{-\gamma/d_w})$ *for some* $\gamma > 0$ *such that* $\zeta_{b,\overline{K}} \sqrt{\log K} = O(K^{\gamma/d_w})$*; (iii) the eigenvalues* G_b *are uniformly bounded from below and above; (iv)* $\int_0^1 \sqrt{1 + \log N_{[]}(w, \mathcal{H}^{\mathrm{R}}, \|\cdot\|_{L^2(Z)})}dw \leq C$ *for some constant* $C > 0$*; (v)* $\|\mathrm{E}_h[(\rho(Y, h(X)) - \rho(Y, h^{\mathrm{R}}(X)))\widetilde{b}^K(W)]\| \leq \|h - h^{\mathrm{R}}\|_\mu$ *for all* $h \in \mathcal{H}^{\mathrm{R}}(\Delta_n)$*; and (vi)* $\widehat{h}^{\mathrm{R}} \in \mathcal{H}^{\mathrm{R}}(\Delta_n)$ *with probability approaching one such that*

$$\max_{1 \leq k \leq K} \mathrm{E} \sup_{h \in \mathcal{H}^{\mathrm{R}}(\Delta_n)} \left|\left(\rho(Y, h(X)) - \rho(Y, h^{\mathrm{R}}(X))\right)\widetilde{b}_k(W)\right|^2 \leq C\Delta_n^{2\kappa}$$

for some $\kappa \in (0, 1]$ *and* $\Delta_n \sum_{K \in \widehat{\mathcal{I}}_n} \sqrt{K/\log\log K} = o_p(1)$.

Assumption 7 (ii) does not impose regularity on the structural function h but rather on the conditional mean function m. Assumption 7 (iv) restricts the complexity of the set of functions $\mathcal{H}^{\mathrm{R}}$ by imposing a finite entropy integral condition. Assumption 7 (v)(vi) were similarly imposed in Chen and Pouzo [2012]. We are now in the position to establish an upper bound for the first and second type error, uniformly of the function class $\mathcal{M}_1(\delta^\circ, \mathsf{r}_n)$, of the data driven test statistic $\mathtt{IT}_n$.

Theorem 4.1. *Let Assumptions 1, 5 (ii) with* J *reeplaced by* K*, and 7 be satisfied. Then, for any* $\varepsilon > 0$ *there exists a constant* $\delta^\circ > 0$ *such that*

$$\limsup_{n \to \infty} \left\{ \mathrm{P}_{h \in \mathcal{H}_0}(\widehat{\mathtt{IT}}_n = 1) + \sup_{h \in \mathcal{M}_1(\delta^\circ, r_n)} \mathrm{P}_h(\widehat{\mathtt{IT}}_n = 0) \right\} \leq \varepsilon, \tag{4.4}$$

where the rate r_n *is given by*

$$\mathsf{r}_n = \left(\sqrt{\log\log n}/n\right)^{2\gamma/(4\gamma + d_w)}.$$

Remark 4.1 (Formulation of Hypotheses)**.** *We see from Theorem 4.1 that the IT rate attains the usual minimax rate of testing within a* $(\log\log n)$ *term and, in particular, does not suffer from the ill-posedness of the underlying inverse problem. This is due to the definition of the class of alternative function* $\mathcal{M}_1(\delta^\circ, \mathsf{r}_n)$*, which measures deviations from the null by the squared norm of the conditional mean function under consideration. Note that it is possible to achieve the same rate of testing as in ST once we are willing to impose* *link conditions between $\|h - h^{\text{R}}\|_\mu$ and $\|m_h(\cdot, h^{\text{R}})\|_{L^2(W)}$. We do not provide this result explicitly in this paper, in the interest of space.*

Remark 4.2 (Comparison of Testing Rates)**.** *From Corollary 3.1 we see that the rate of testing suffers only in the case of ST from the ill-posed inverse problem. Nevertheless, in the mildly ill-posed case, the rate of ST can be even better than of IT, that is,*

$$n^{-2p/(4(p+\zeta)+d_x)} < n^{-2\gamma/(4\gamma+d_w)}$$

if and only if the dimension of W satisfies

$$d_w > \frac{\gamma}{p}\big(4\zeta + d_x\big).$$

In contrast the rate in case of ST is always slower than the IT rate in the severely ill-posed case.

5. Empirical Applications, Further Simulations Studies

5.1. Empirical Applications

We present three empirical applications of our adaptive structure-space tests for NPIV models. The first one tests for monotonicity of demand for differentiated products in IO. The second one tests for monotonicity of gasoline demand function. The third one tests for monotonicity or parametric specification of Engel curves for non-durable good consumption.

In all of the empirical applications, we implement the adaptive test $\widehat{\texttt{ST}}_n$ given in (2.2), using the adaptive critical values $\widehat{\eta}_J(\alpha)$ presented in Remark 3.1 (see Appendix C for bootstrap critical values). Throughout this section, we use quadratic B-spline basis with varying number of knots to approximate h and set $K(J) = 2J$. For any sieve dimension $J \in \widehat{\mathcal{I}}_n$ (the ES index set), we denote the corresponding "standardized test value" as

$$\widehat{\mathcal{W}}_J := \frac{n\,\widehat{D}_J(\widehat{h}_J^{\text{R}})}{\sqrt{2}\widehat{\eta}_J(\alpha)\,\widehat{\text{v}}_J(\widehat{h}_J^{\text{R}})}$$

at the nominal level $\alpha = 0.05$. The null hypothesis is rejected whenever $\widehat{\mathcal{W}}_J > 1$ for some $J \in \widehat{\mathcal{I}}_n$. Tables below reports a set $\widehat{\mathcal{J}} \subset \widehat{\mathcal{I}}_n$, which equals to $\arg\max_{J\in\widehat{\mathcal{I}}_n} \widehat{\mathcal{W}}_J$ when the null is not rejected at the nominal level $\alpha = 0.05$; and equals to $\{J \in \widehat{\mathcal{I}}_n : \widehat{\mathcal{W}}_J > 1\}$ when the null is rejected at the nominal level $\alpha = 0.05$. Let $\widehat{J}$ denote the minimal integer of $\widehat{\mathcal{J}}$. In all the tables below, we report $\widehat{\mathcal{W}}_{\widehat{J}}$, and its corresponding p value, which should, by Bonferroni correction, be compared to the nominal level 0.05 divided by the cardinality of $\widehat{\mathcal{I}}_n$, which differ in the applications presented below.

5.1.1. Adaptive ST testing for a multi-Product Demand systems

Recently Compiani [2019] applies nonparametric estimation of a NPIV model as an alternative to BLP semiparametric specification. He directly imposes shape restrictions to reduce dimensionality in nonparametric estimation of own-price elasticity and cross-price elasticity of multi-product demand. Following Compiani [2019] we use the Nielsen scanner data set on demand for strawberries in California.

Our application here is to provide adaptive nonparametric hypothesis testing on the monotonicity of demand in its own price. First notice that by imposing an index restriction and a connected substitute assumption the empirical model can be written in NPIV form, following Berry and Haile [2014] and Compiani [2019]:

$$Y_o = h(P_o, P_{no}, P_{other}, S_o, In) - U, \qquad \mathrm{E}[U|W_o, W_{no}, W_{other}, W_{S_o}, In] = 0,$$

where Y_o denotes a measure of taste for organic products, S_o denotes the share of the organic products, In income, and U unobserved shocks for organic produce. The vector (P_o, P_{no}, P_{other}) denotes the prices of organic strawberries, non-organic strawberries, and non-strawberry fresh fruit, respectively. To account for possible endogeneity of prices we follow Compiani [2019] and use Hausman-type instrumental variables denoted by (W_o, W_{no}, W_{other}) and shipping-point spot prices W_{S_o} as a proxy for the wholesale prices faced by retailers.

Income	H_0: h is decreasing in P_o				H_0: h is increasing in P_o			
groups	$\widehat{\mathcal{W}}_{\widehat{\mathcal{J}}}$	p val.	reject H_0?	$\widehat{\mathcal{J}}$	$\widehat{\mathcal{W}}_{\widehat{\mathcal{J}}}$	p val.	reject H_0?	$\widehat{\mathcal{J}}$
low	0.552	0.103	no	$\{8\}$	1.852	0.003	yes	$\{6\}$
middle	0.234	0.246	no	$\{10\}$	6.720	0.000	yes	$\{6, 8, 10\}$
high	1.526	0.000	yes	$\{8, 10, 12\}$	8.883	0.000	yes	$\{6, 8, 10, 12\}$

Table 2: Adaptive testing for monotonic demand for organic products.

Data on prices and quantities come from the 2014 Nielsen scanner data set. For each market, the most granular unit of observation in the Nielsen data is a UPC (i.e. a specific bar code). We consider a "low income", "middle income", and "high income" group based on individuals between the 5%–25%, 40%–60%, and 75%–95% quantile, respectively, of the income distribution. The resulting sample sizes for the three subgroups are 1509, 1491, and 2093. We implement our adaptive test $\widehat{\mathtt{ST}}_n$ by making use of a semiparametric specification of the structural demand function: we consider the tensor product of quadratic B-splines $q^{J_1}(P_o)$ and the vector $(1, In)$, hence $J = 2J_1$. The other variables are accounted for separately, that is, P_{no} and P_{other} fixed with two interior knots and market shares S_o linearly. The cardinality of the index set $\widehat{\mathcal{I}}_n$ changes for the different subgroups considered:

We have $\widehat{\mathcal{I}}_n = \{6, 8, 10\}$ for the low and middle income groups but $\widehat{\mathcal{I}}_n = \{6, 8, 10, 12\}$ for the high income group.

According to Table 2, our adaptive test rejects the null of weakly decreasing demand curve (in own price) at the nominal level $\alpha = 0.05$ for the high income group, but fails to reject the decreasing demand for the low and middle income groups. In addition, our adaptive test rejects the null of weakly increasing demand curve (in own price) for all income groups.

5.1.2. Adaptive testing for monotonicity in gasoline demand

In this subsection, we revisit the problem whether the gasoline demand is monotone decreasing in its own price or not. We consider the following partially linear specification of the demand function:

$$Y = h(P, In) + \gamma' H + U, \qquad \mathrm{E}[U|W, In, H] = 0$$

where Y denotes annual log-gasoline consumption of a household, P log-price of gasoline (average local price), In log-household income, H are control variables (such as log age of the household respondent, the log household size, the log number of drivers, and the number of workers in the household), and W distance to major oil platform as instrumental variables. We allow for price P to be endogenous, but assume that (In, H) is exogenous.

Income	H_0: h is decreasing in P				H_0: h is increasing in P			
groups	$\widehat{\mathcal{W}}_{\widehat{\mathcal{J}}}$	p val.	reject H_0?	$\widehat{\mathcal{J}}$	$\widehat{\mathcal{W}}_{\widehat{\mathcal{J}}}$	p val.	reject H_0?	$\widehat{\mathcal{J}}$
low	0.433	0.111	no	$\{8\}$	1.268	0.000	yes	$\{6, 8\}$
high	11.703	0.000	yes	$\{8, 10\}$	1.625	0.000	yes	$\{6, 8, 10\}$

Table 3: Adaptive testing for monotonicity of gasoline demand.

Consumer theory requires downward-sloping compensated demand curves. In the following we provide a test for monotonicity of the uncompensated (Marshallian) demand curves. The data are from the 2001 National Household Travel Survey (NHTS), which surveys, on a household-level, the civilian noninstitutionalized population in the United States. Following Blundell et al. [2017] we restrict the analysis to households with a white respondent, two or more adults, at least one child under age 16, and at least one driver. The resulting sample contains 3,640 observations.[9]

Chetverikov and Wilhelm [2017] and Freyberger and Reeves [2019] have used this data set to estimate the gasoline demand by assuming a decreasing demand curve. Instead we

[9]We thank Matthias Parey for sharing the dataset with us. We refer the reader to Blundell et al. [2017] for a detailed description of the data.

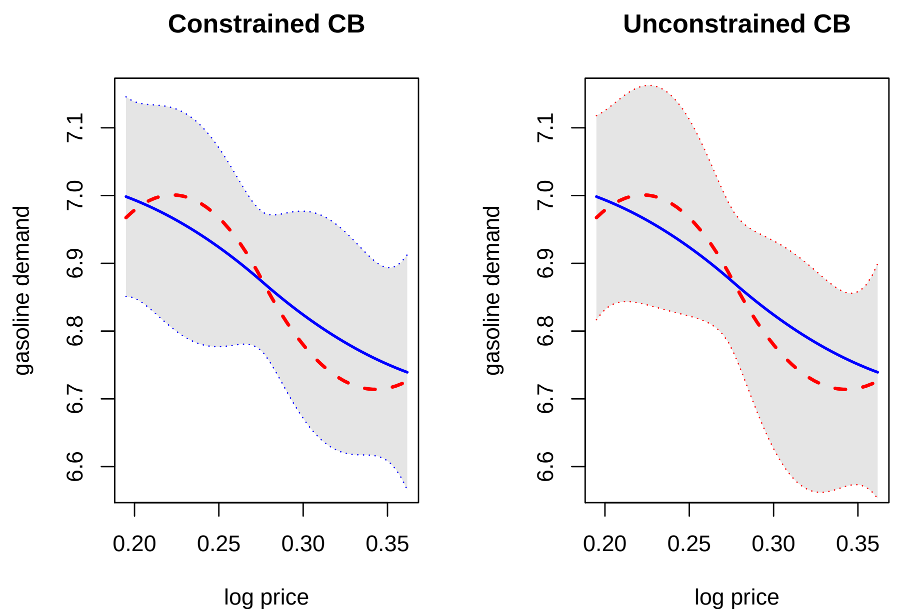


Figure 4: Estimated demand curves for the low income group using $\widehat{J} = 8$: constraint estimator (blue solid line) and unconstraint estimator (red dashed line). LHS: Dotted blue lines show 95% uniform CBs based on constrained estimator. RHS: Dotted red lines show 95% uniform CBs based on unconstrained estimator. All other variables are fixed to their median levels (implying $32,500 of income).

aim at testing the monotonicity of the gasoline demand here. We consider a semiparametric specification similar to theirs to approximate the unknown function h, that is, our test is based on the tensor product of quadratic B-splines $q^{J_1}(P)$ and the vector $(1, In)$, hence $J = 2J_1$. To analyze heterogeneity in demand for different income levels we make use of two different subsamples of the data set. We consider a sample of $n = 803$ "low income" households whose income is below the first quartile, and a sample of $n = 1369$ "high income" households whose income is weakly above the third quartile. When computing the ES index set we obtain $\widehat{\mathcal{I}}_n = \{6, 8\}$ for low income group and $\widehat{\mathcal{I}}_n = \{6, 8, 10\}$ for the high income group.

According to Table 3, our adaptive test rejects the null of weakly decreasing gasoline demand at the nominal level $\alpha = 0.05$ for the high income group, but fails to reject the decreasing demand for the low income group. In addition, our adaptive test rejects the null of weakly increasing gasoline demand curve for both income groups.

We illustrate our testing results in Figures 4 and 5 which shows the graphs of restricted and unrestricted NPIV estimators for the low and high income groups, respectively. The estimators are implemented using the choice of the sieve dimension given by $\widehat{J} = 8$ in both groups. Variables other than price are fixed to their median level at each subgroup. In

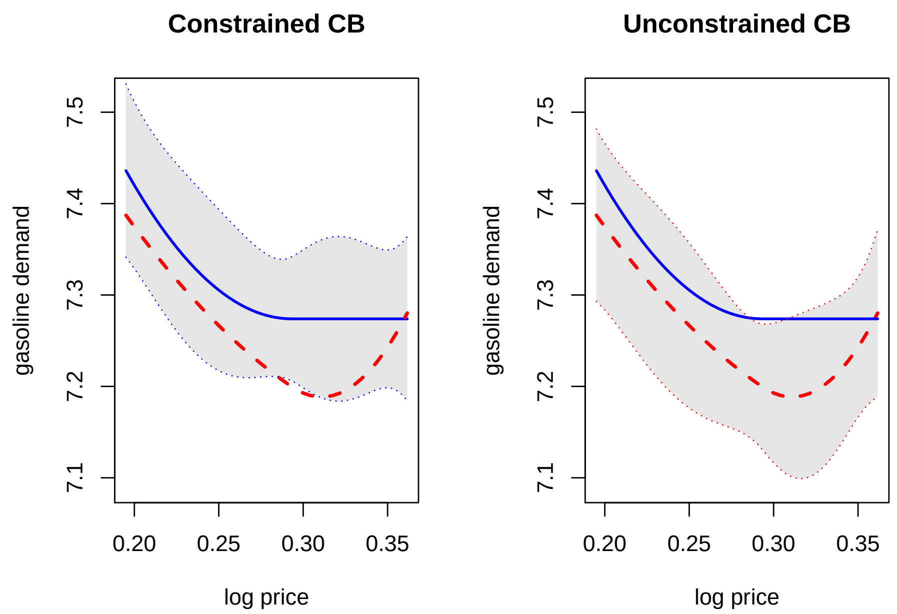


Figure 5: Estimated demand curves for the high income group using $\widehat{J} = 8$: constraint estimator (blue solid line) and unconstraint estimator (red dashed line). LHS: Dotted blue lines show 95% uniform CBs based on constrained estimator. RHS: Dotted red lines show 95% uniform CBs based on unconstrained estimator. All other variables are fixed to their median levels (implying $120,000 of income).

particular, the median income is given by roughly $32,500 and $120,000 for the two income groups considered. We also provide 95% bootstrap uniform confidence bands (CBs), using 1000 bootstrap iterations. Both figures are in line with our adaptive testing results reported in Table 3, that is, only the demand curves for high income households seem to violate a montonically decreasing shape. For high income households the unrestricted demand curves are slightly outside of the 95% confidence bands of the restricted NPIV estimator.

5.1.3. Adaptive testing for Engel Curves

Engel curves play a central role in the analysis of consumer demand and describe the household budget share Y_ℓ for non-durable goods ℓ as a function of household log-total expenditure X. Blundell et al. [2007] estimated a system of nonparametric Engel curves as functions of endogenous log-total expenditure and family size, using log-gross earnings of the head of household as an instrument W. We use a subset of their data from the 1995 British Family Expenditure Survey, with the head of household aged between 20 and 55 and in work, and household with one or two children. This leaves a sample of size $n = 1027$. As an illustration we consider Engel curves $h_\ell(X)$ for four non-durable goods ℓ: "food in", "fuel", "travel", and "leisure": $E[Y_j - h_\ell(X)|W] = 0$. We use the same quadratic

B-spline basis with up to 3 knots to approximate all the Engel curves. Hence the index set $\widehat{\mathcal{I}}_n = \{3, 4, 5, 6\}$ is the same for the different Engel curves. Table 4 reports the adaptive test for weak monotonicity of Engel curves.

	H_0: h is increasing				H_0: h is decreasing			
Goods	$\widehat{\mathcal{W}}_{\widehat{J}}$	p value	reject H_0?	$\widehat{J}$	$\widehat{\mathcal{W}}_{\widehat{J}}$	p value	reject H_0?	$\widehat{J}$
"food in"	1.347	0.002	yes	$\{3\}$	-0.269	0.821	no	$\{3\}$
"fuel"	4.484	0.000	yes	$\{3, 4\}$	0.347	0.114	no	$\{3\}$
"travel"	2.074	0.000	yes	$\{3\}$	0.255	0.155	no	$\{3\}$
"leisure"	0.295	0.151	no	$\{6\}$	2.550	0.000	yes	$\{3, 4\}$

Table 4: Adaptive testing for monotonicity of Engel curves.

Table 4 reports that our test rejects increasing Engel curves for "food in", "fuel", and "travel" categories, and also rejects decreasing Engel curve for "leisure" at the 0.05 nominal level. Previously, to decide whether the Engel curves is strictly monotonic, estimated derivatives of these function together with their 95% uniform confidence bands were also provided in Chen and Christensen [2018, Figure 4]. Those confidence bands contain zero almost over the whole support of household expenditure. It is interesting to see that our adaptive test is more informative about monotonicity in certain directions that are not obvious from their uniform confidence bands.

	H_0: h is linear				H_0: h is quadratic			
Goods	$\widehat{\mathcal{W}}_{\widehat{J}}$	p value	reject H_0?	$\widehat{J}$	$\widehat{\mathcal{W}}_{\widehat{J}}$	p value	reject H_0?	$\widehat{J}$
"food in"	-0.169	0.644	no	$\{3\}$	0.230	0.186	no	$\{3\}$
"fuel"	1.174	0.004	yes	$\{3\}$	-0.089	0.502	no	$\{6\}$
"travel"	1.052	0.007	yes	$\{3\}$	-0.108	0.531	no	$\{5\}$
"leisure"	0.536	0.074	no	$\{6\}$	0.454	0.098	no	$\{6\}$

Table 5: Adaptive testing for linear/quadratic specification of Engel curves.

The most popular class of parametric demand systems is the almost ideal class, pioneered by Deaton and Muellbauer [1980], where budget shares are assumed to be linear in log-total expenditure. Banks et al. [1997] proposed a popular extension of this model to include a squared term in log-total expenditure. Table 5 presents tests for either a linear or quadratic specification of the Engel curves for the four goods considered above. According to this table, at the nominal level $\alpha = 0.05$, our adaptive ST fails to reject quadratic form for all the goods, while rejects a linear Engel curve for fuel and travel goods.

5.2. A Monte Carlo Study: Testing for Parametric Restrictions

In this subsection, we compare the finite sample performances of the adaptive structure-space test ($\widehat{\mathtt{ST}}_n$) and the adaptive image-space test ($\widehat{\mathtt{IT}}_n$) when h takes a parametric form under the null. The results are based on 5000 Monte Carlo replications for every experiment.

We consider two different designs similar to the simulation set up in (2.7). First, we set $X_i = \Phi(X_i^*)$ and $W_i = \Phi(W_i^*)$, where the random vector (X_i^*, W_i^*, U_i) is generated according to

$$\begin{pmatrix} X_i^* \\ W_i^* \\ U_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \xi & 0.3 \\ \xi & 1 & 0 \\ 0.3 & 0 & \sigma_U^2 \end{pmatrix} \right) . \tag{5.1}$$

Second, we consider multivariate conditioning variable $W = (W_1, W_2)$. To do so, we set $X_i = \Phi(X_i^*)$, $W_{1i} = \Phi(W_{1i}^*)$, and $W_{2i} = \Phi(W_{2i}^*)$, where

$$\begin{pmatrix} X_i^* \\ W_{1i}^* \\ W_{2i}^* \\ U_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \xi & 0.4 & 0.3 \\ \xi & 1 & 0 & 0 \\ 0.4 & 0 & 1 & 0 \\ 0.3 & 0 & 0 & \sigma_U^2 \end{pmatrix} \right) . \tag{5.2}$$

In both experiment designs, (5.1) and (5.2), we let $\sigma_U = 0.2$, and Y_i be generated according to (2.6) with the structural function

$$h(x) = -x/5 + c_A x^2 + (c_B x)^3,$$

where the constants c_A and c_B will vary in the experiments below. Specifically, we consider either a linear (i.e., $(c_A, c_B) = (0, 0)$) or quadratic form (i.e., $(c_A, c_B) = (1, 0)$) as the null hypothesis. We evaluate the power of our tests to detect either quadratic models (i.e., $(c_A, c_B) = (1, 0)$) or cubic ones (i.e., $(c_A, c_B) = (1, 1)$ when the null is linear or $(c_A, c_B) = (1, 5)$ when the null is quadratic).

The simulation sample size is $n = 500$. We implement the adaptive structure-space test $\widehat{\mathtt{ST}}_n$ given in (2.2) using quadratic B-spline basis functions with varying number of knots. The number of knots is varied within the index set $\widehat{\mathcal{I}}_n$ as implemented in the last subsection, also with $K(J) = 2J$. In addition, we implement the adaptive image-space test $\widehat{\mathtt{IT}}_n$ given in (4.2) with using quadratic B-spline basis functions with varying number of knots, where h_0 is replaced by the 2SLS estimator.

In Table 6, we depict the empirical rejection probabilities of the test statistics $\widehat{\mathtt{ST}}_n$ and $\widehat{\mathtt{IT}}_n$, and their bootstrapped versions $\widehat{\mathtt{ST}}_n^{\mathtt{B}}$ and $\widehat{\mathtt{IT}}_n^{\mathtt{B}}$, with 200 bootstrap replications for the bootstrapped critical values. Again, the adaptive tests and their respectively bootstrapped

Design	ξ	*Null* (c_A, c_B)	*Alt.* (c_A, c_B)	*Empirical Rejection prob./Average dimension choice* $\widehat{\mathtt{ST}}_n^{\mathtt{B}}$	$\widehat{\mathtt{ST}}_n$	$\widehat{J}_{\mathtt{ST}}$	$\widehat{\mathtt{IT}}_n^{\mathtt{B}}$	$\widehat{\mathtt{IT}}_n$	$\widehat{K}_{\mathtt{IT}}$	t-test
(5.1)	0.5	$(0,0)$		0.046	0.055	3.92	0.051	0.047	3.81	0.026
$d_x = d_w$		$(1,0)$		0.022	0.031	4.17	0.029	0.024	4.14	0.001
		$(0,0)$	$(1,0)$	0.072	0.088	3.78	0.084	0.080	3.76	0.084
		$(0,0)$	$(1,1)$	0.272	0.279	3.40	0.252	0.245	3.54	0.380
		$(1,0)$	$(1,5)$	0.092	0.120	4.16	0.038	0.037	4.12	0.087
	0.7	$(0,0)$		0.046	0.049	4.13	0.042	0.049	3.82	0.045
		$(1,0)$		0.026	0.029	4.53	0.023	0.026	4.12	0.027
		$(0,0)$	$(1,0)$	0.185	0.167	3.68	0.184	0.198	3.57	0.293
		$(0,0)$	$(1,1)$	0.810	0.815	3.06	0.802	0.822	3.14	0.912
		$(1,0)$	$(1,5)$	0.515	0.538	4.08	0.295	0.330	4.12	0.728
(5.2)	0.5	$(0,0)$		0.052	0.052	3.98	0.041	0.039	7.82	0.035
$d_x < d_w$		$(1,0)$		0.030	0.032	4.20	0.028	0.032	8.12	0.007
		$(0,0)$	$(1,0)$	0.111	0.115	3.79	0.061	0.052	7.73	0.121
		$(0,0)$	$(1,1)$	0.430	0.461	3.31	0.113	0.116	7.42	0.536
		$(1,0)$	$(1,5)$	0.321	0.340	4.28	0.031	0.032	8.17	0.255
	0.7	$(0,0)$		0.042	0.050	4.10	0.036	0.040	7.84	0.045
		$(1,0)$		0.026	0.031	4.43	0.033	0.034	8.13	0.034
		$(0,0)$	$(1,0)$	0.209	0.206	3.61	0.064	0.063	7.46	0.322
		$(0,0)$	$(1,1)$	0.874	0.883	3.04	0.370	0.354	6.54	0.942
		$(1,0)$	$(1,5)$	0.687	0.700	4.08	0.054	0.052	8.25	0.797

Table 6: Empirical Rejection probabilities of $\widehat{\mathtt{ST}}_n^{\mathtt{B}}$, $\widehat{\mathtt{ST}}_n$, $\widehat{\mathtt{IT}}_n^{\mathtt{B}}$, $\widehat{\mathtt{IT}}_n$-tests, and t-test (of the hypothesis $(c_A, c_B) = (0, 0)$ if null is linear and of $(c_A, c_B) = (1, 0)$ if null is quadratic) under 5% nominal level, $n = 500$.

versions perform similarly in terms of sizes and powers. All these adaptive tests are compared with the asymptotic t-test of the hypothesis $c_A = 0$ if the null is linear and of $c_B = 0$ if the null is quadratic. We see that the adaptive tests are overall not very conservative in comparison to the t-test that imposes specific parametric restrictions. The powers of all the tests increases as ξ increases (i.e., the instrument gets stronger). Very remarkably, the adaptive ST performs as well as the asymptotic t-test does. For the first experiment design given in (5.1) (with $d_x = d_w$), we see that the adaptive ST is somewhat more powerful than the adaptive IT when the true model is cubic, i.e., $(c_A, c_B) = (1, 5)$. But, in the second simulation design (5.2) (with $d_x < d_w$), the adaptive ST is much more powerful than the adaptive IT. This finding is consistent with our theoretical findings as described in Remark 4.2. This severe power difference also holds when cubic or quartic B-splines are used as the vector of instrument basis functions $b^K(W)$.

6. Conclusion

In this paper, we propose data-driven, minimax rate-adaptive hypothesis testing on a structural function in semi-nonparametric conditional moment restrictions. Our main focus is the adaptive structure-space test (ST) that is based on a leave-one-out sieve estimate of quadratic distance between the structural functions of endogenous variables in NPIV models. But we also present the minimax rate-adaptive image-space test (IT) as comparison. This is because our sieve IT is related to the Bierens' type tests that have been utilized by many existing papers on inference for semi-nonparametric conditional moment restrictions. All the prior existing papers on testing for NPIV models do not consider data-driven choice of tuning parameters, however. For both tests, we first establish their respective minimax rate of testing against classes of nonparametric alternative models. We then provide computationally simple data-driven choices of sieve tuning parameters and adaptive critical values. The resulting tests attain the optimal minimax rate of testing, adapt to the unknown smoothness of functions, and are robust to the unknown degree of ill-posedness (endogeneity). Data-driven confidence sets are easily obtained by inverting the adaptive ST. Monte Carlo studies and empirical applications demonstrate that our simple, adaptive ST has good size and power properties in finite samples for testing monotonicity or parametric restrictions in NPIV models, without the need of using computationally intensive bootstrapped critical values.

References

D. Al Mohamad, E. W. Van Zwet, E. Cator, and J. J. Goeman. Adaptive critical value for constrained likelihood ratio testing. *arXiv preprint arXiv:1806.01325*, 2018.

J. Banks, R. Blundell, and A. Lewbel. Quadratic engel curves and consumer demand. *Review of Economics and Statistics*, 79(4):527–539, 1997.

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

H. J. Bierens. A consistent conditional moment test of functional form. *Econometrica*, 58 (6):1443–1458, 1990.

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

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

C. Breunig. Goodness-of-fit tests based on series estimators in nonparametric instrumental regression. *Journal of Econometrics*, 184(2):328–346, 2015.

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

C. Butucea, C. Matias, and C. Pouet. Adaptive goodness-of-fit testing from indirect observations. *Ann. Inst. Henri Poincaré Probab. Stat*, 45(2):352–372, 2009.

T. T. Cai and M. G. Low. Adaptive confidence balls. *The Annals of Statistics*, 34(1): 202–228, 2006.

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

X. Chen and T. Christensen. Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation, 2015.

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

X. Chen and D. Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 01 2012.

X. Chen and D. Pouzo. Sieve quasi likelihood ratio inference on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–1079, 2015.

X. Chen and M. Reiß. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27(03):497–521, 2011.

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

V. Chernozhukov, W. K. Newey, and A. Santos. Constrained conditional moment restriction models. *arXiv preprint arXiv:1509.06311*, 2015.

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

O. Collier, L. Comminges, and A. B. Tsybakov. Minimax estimation of linear and quadratic functionals on sparsity classes. *The Annals of Statistics*, 45(3):923–958, 2017.

G. Compiani. Market counterfactuals and the specification of multi-product demand: A nonparametric approach. *Available at SSRN 3134152*, 2019.

A. Deaton and J. Muellbauer. An almost ideal demand system. *American Economic Review*, pages 312–326, 1980.

Z. Fang and J. Seo. A general framework for inference on shape restrictions. *arXiv preprint arXiv:1910.07689*, 2019.

J. Freyberger and B. Reeves. Inference under shape restrictions. Working paper, University of Wisconsin, 2019.

E. Gautier and E. Le Pennec. Adaptive estimation in the nonparametric random coefficients binary choice model by needlet thresholding. *Electronic Journal of Statistics*, 12(1):277–320, 2018.

E. Gine and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2016.

E. Guerre and P. Lavergne. Data-driven rate-optimal specification testing in regression models. *The Annals of Statistics*, 33(2):840–870, 2005.

L. P. Hansen and T. J. Sargent. *Robustness*. Princeton university press, 2008.

J. L. Horowitz. Testing a parametric model against a nonparametric alternative with identification through instrumental variables. *Econometrica*, 74(2):521–538, 2006.

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

J. L. Horowitz and V. G. Spokoiny. An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica*, 69(3):599–631, 2001.

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

Y. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Math. Methods Statist*, 2(2):85–114, 1993.

Y. I. Ingster, T. Sapatinas, I. A. Suslina, et al. Minimax signal detection in ill-posed inverse problems. *The Annals of Statistics*, 40(3):1524–1549, 2012.

M. Jansson and D. Pouzo. Towards a general large sample theory for regularized estimators. Working paper, University of California, Berkeley, 2019.

J. Lei. Adaptive global testing for functional linear models. *Journal of the American Statistical Association*, 109(506):624–634, 2014.

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

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

J. Robins and A. Van Der Vaart. Adaptive nonparametric confidence sets. *The Annals of Statistics*, 34(1):229–253, 2006.

A. Santos. Inference in nonparametric instrumental variables with partial identification. *Econometrica*, 80(1):213–275, 2012.

M. J. Silvapulle and P. K. Sen. *Constrained statistical inference: Inequality, order and shape restrictions*. John Wiley & Sons, 2005.

V. G. Spokoiny. Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, 24 (6):2477–2498, 1996.

J. Tao. Inference for point and partially identified semi-nonparametric conditional moment models. In *Unpublished Manuscript, University of Wisconsin-Madison*, 2014.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009.

Y. Zhu. Inference in nonparametric/semiparametric moment equality models with shape restrictions. *Quantitative Economics*, 11(2):609–636, 2020.

A. Proofs of Minimax ST Results in Subsection 3.2

We introduce additional notation used in the proofs. For a $r \times c$ matrix M with $r \leq c$ and full row rank r we let M_l^- denote its left pseudoinverse, namely $(M'M)^-M'$ where $'$ denotes transpose and $-$ denotes generalized inverse. We define $A = [S'G_b^{-1}S]^{-1}S'G_b^{-1}$ where $S = \mathrm{E}[b^K(W)\psi^J(X)']$. For all $J \geq 1$ such that $s_J = s_{\min}(G_b^{-1/2}SG^{-1/2}) > 0$ it holds

$$\left\|G^{1/2}AG_b^{1/2}\right\| = \left\|G^{1/2}[S'G_b^{-1}S]^{-1}S'G_b^{-1/2}\right\| = \left\|[G^{-1/2}S'G_b^{-1}SG^{-1/2}]^{-1}G^{-1/2}S'G_b^{-1/2}\right\|$$
$$= \left\| \left(G_b^{-1/2}SG^{-1/2}\right)_l^- \right\| = s_J^{-1}.$$

Below, we make use of the notation $V_i^J := (Y_i - h_0(X_i))G^{1/2}Ab^K(W_i)$. We also introduce the projection $Q_J h(\cdot) = \psi^J(\cdot)'A\,\mathrm{E}[h(X)b^K(W)]$ for $h \in L^2_\mu$. We introduce the normalization term

$$\mathrm{v}_J(h_0) = \big\| \mathrm{E}_h[\big(Y - h_0(X)\big)^2 G^{1/2} A\, b^{K(J)}(W) b^{K(J)}(W)' A' G^{1/2}] \big\|_F.$$

For simplicity of notation, we write $\widehat{D}_J$ instead of $\widehat{D}_J(h_0)$ and v_J instead of $\mathrm{v}_J(h_0)$.

Proof of Theorem 3.1. For the proof of the result, we proceed in three steps. First, we bound the first type error of the test statistic

$$\widetilde{\mathtt{ST}}_{n,J} = \mathbb{1}\left\{ n\widehat{D}_J > \eta' \mathrm{v}_J \right\}$$

for some constant $\eta' > 0$. Second, we bound the second type error of $\widetilde{\mathtt{ST}}_{n,J}$ where η' is replaced by $\eta'' > 0$. Third, we make the dependence of η',η'' with η explicit and show that the derived bounds in the previous steps are sufficient to control the first and second type error of our test statistic $\widetilde{\mathtt{ST}}_{n,J}$.

Step 1: Under the simple null hypothesis we have $\|h - h_0\|_\mu = 0$ and hence we obtain for any $\varepsilon > 0$ and η sufficiently large that

$$\mathrm{P}_{h_0}(\widetilde{\mathtt{ST}}_{n,J} = 1) = \mathrm{P}_{h_0}\left(n\widehat{D}_J > \eta' \,\mathrm{v}_J \right) \le \varepsilon$$

by the upper bound derived in (F.4).

Step 2: From the Step 2. of the proof of Theorem 3.3 (with J^* replaced by J and using that η''_{J^*} replaced by η'') we obtain uniformly over $h \in \mathcal{H}_1(\delta^*, r_{n,J})$ that

$$\mathrm{P}_h(\widetilde{\mathtt{ST}}_{n,J} = 0) = \mathrm{P}_h\left(\widehat{D}_J \le \frac{\eta'' \,\mathrm{v}_J}{n} \right) = o(1).$$

Step 3: Finally, we account for estimation of the normalization factor v_J. Now since $|\widehat{\mathrm{v}}_J - \mathrm{v}_J| \le c_0 \mathrm{v}_J$ wpa1 for some $0 < c_0 < 1$ by Lemma F.6 it holds $(1-c_0)\mathrm{v}_J \le \widehat{\mathrm{v}}_J \le (1+c_0)\mathrm{v}_J$ wpa1. Let η' and η'' be such that

$$\sqrt{2}\eta = \frac{\eta'}{(1-c_0)} = \frac{\eta''}{(1+c_0)}.$$

Hence, we control the first type error of the test $\widetilde{\mathtt{ST}}_{n,J}$ as follows

$$\begin{aligned}
\mathrm{P}_{h_0}\left(\mathtt{ST}_{n,J} = 1\right) &\le \mathrm{P}_{h_0}\left(\mathtt{ST}_{n,J} = 1,\, \widehat{\mathrm{v}}_J \ge (1-c_0)\mathrm{v}_J\right) + \mathrm{P}_{h_0}\left(\widehat{\mathrm{v}}_J < (1-c_0)\mathrm{v}_J\right) \\
&\le \mathrm{P}_{h_0}\left(\widehat{D}_J > \sqrt{2}\eta \mathrm{v}_J(1-c_0) \right) + o(1) \\
&= \mathrm{P}_{h_0}\left(\widehat{D}_J > \eta' \mathrm{v}_J \right) + o(1) \\
&= o(1)
\end{aligned}$$

where the last equation is due to Step 1. To bound the second error term of $\mathtt{ST}_{n,J}$ we calculate uniformly over $h \in \mathcal{H}_1(\delta^*, r_{n,J})$ that

$$\begin{aligned}
\mathrm{P}_h\left(\widetilde{\mathtt{ST}}_{n,J} = 0\right) &\leq \mathrm{P}_h\left(\widetilde{\mathtt{ST}}_{n,J} = 0, \quad \widehat{\mathrm{v}}_J \leq (1+c_0)\mathrm{v}_J\right) + \mathrm{P}_{h_0}\left(\widehat{\mathrm{v}}_J > (1+c_0)\mathrm{v}_J\right) \\
&= \mathrm{P}_h\left(|\widehat{D}_J| \leq \eta''\mathrm{v}_J\right) + o(1) \\
&= o(1),
\end{aligned}$$

where the last equation is due to Step 2. □

Proof of Corollary 3.1. We make use of the observation $s_J^{-1} = (1+o(1))\tau_J$.

Proof of (3.3). The choice of $J \sim n^{2d_x/(4(p+\zeta)+d_x)}$ implies

$$n^{-1}\tau_J^2\sqrt{J} \sim n^{-1}J^{1/2+2\zeta/d_x} \sim n^{-4p/(4(p+\zeta)+d_x)}$$

and for the bias term $J^{-2p/d_x} \sim n^{-4p/(4(p+\zeta)+d_x)}$.

Proof of (3.4). The choice of $J \sim (\log n - \zeta^{-1}(2p+d_x)\log\log n)^{d_x/\zeta}$ implies

$$\begin{aligned}
n^{-1}\tau_J^2\sqrt{J} \sim n^{-1}\sqrt{J}\exp(J^{\zeta/d_x}) &\sim \Big(\log n - \frac{2p+d_x}{\zeta}\log\log n\Big)^{d_x/(2\zeta)}(\log n)^{-(2p+1)/\zeta} \\
&\lesssim (\log n)^{d_x/(2\zeta)}(\log n)^{-(2p+d_x)/\zeta} \\
&\lesssim (\log n)^{-2p/\zeta}
\end{aligned}$$

and for the bias term $J^{-2p/d_x} \sim (\log n)^{-2p/\zeta}$, which yields equation (3.4). □

Proof of Theorem 3.2. It is sufficient to consider the Gaussian reduced-form NPIR as in Chen and Reiß [2011]. From the proof of Collier et al. [2017, Lemma 3] we infer for any $h_\theta \in \mathcal{H}_1(\delta_*, r_n)$ that

$$\begin{aligned}
\inf_{\mathtt{T}_n}\Big\{\mathrm{P}_{h_0}(\mathtt{T}_n = 1) + \sup_{h \in \mathcal{H}_1(\delta_*, r_n)}\mathrm{P}_h(\mathtt{T}_n = 0)\Big\} &\geq \inf_{\mathtt{T}_n}\Big\{\mathrm{P}_{h_0}(\mathtt{T}_n = 1) + \mathrm{P}_{h_\theta}(\mathtt{T}_n = 0)\Big\} \\
&\geq 1 - \mathcal{V}(\mathrm{P}_{h_\theta}, \mathrm{P}_{h_0}) \\
&\geq 1 - \sqrt{\chi^2(\mathrm{P}_{h_\theta}, \mathrm{P}_{h_0})}, \qquad \text{(A.1)}
\end{aligned}$$

where $\mathcal{V}(\cdot,\cdot)$ denotes the total variation distance and $\chi^2(\cdot,\cdot)$ denotes the χ^2 divergence. We consider $\theta = (\theta_j)_{j\geq 1}$ with $\theta_j \in \{-1, 1\}$ and introduce the test function

$$h_\theta = h_0 + \sqrt{\frac{\delta_*}{n}}\sum_{j=1}^{J}\tau_j^2\theta_j\widetilde{\psi}_j\Big(\sum_{j=1}^{J}\tau_j^4\Big)^{-1/4},$$

where $\{\widetilde{\psi}_j\}_{j\geq 1}$ forms an orthonormal basis in L^2_μ and the dimension parameter J satisfies the inequality restriction

$$\frac{\delta_*}{n}\sqrt{\sum_{j=1}^{J}\tau_j^4 j^{4p}} \leq C_0 \tag{A.2}$$

for some constant $C_0 > 0$. Therefore, orthonormality of the basis functions $\{\widetilde{\psi}_j\}_{j\geq 1}$ in L^2_μ together with the Cauchy-Schwarz inequality implies

$$\sum_{j=1}^{\infty}\langle h_\theta - h_0, \widetilde{\psi}_j\rangle_\mu^2 j^{2p} = \frac{\delta_*}{n}\sum_{j=1}^{J}\tau_j^4 j^{2p}\Big(\sum_{l=1}^{J}\tau_l^4\Big)^{-1/2} \leq \frac{\delta_*}{n}\sqrt{\sum_{j=1}^{J}\tau_j^4 j^{4p}} \leq C_0$$

and we conclude that $h_\theta - h_0$ attains the sieve approximation error imposed in Assumption 2. Denoting $r_n = n^{-1/2}\big(\sum_{j=1}^{J}\tau_j^4\big)^{1/4}$ we obtain

$$\|h_\theta - h_0\|_\mu^2 = \frac{\delta_*}{n}\Big(\sum_{j=1}^{J}\tau_j^4\Big)^{1/2}\int \widetilde{\psi}_j^2(x)\mu(x)dx = \frac{\delta_*}{n}\Big(\sum_{j=1}^{J}\tau_j^4\Big)^{1/2} = \delta^* r_n^2$$

and hence, $h_\theta \in \mathcal{H}_1(\delta^*, r_n)$. Under P_{h_θ} the conditional distribution of Y given W is Gaussian with mean $Th_\theta(W)$ and variance 1. We may assume that $\{\lambda_j, \widetilde{\psi}_j, \widetilde{b}_j\}$, $j \geq 1$, forms a singular value decomposition of the conditional expectation operator T. The total variation of P_{h_θ} and P_{h_0} thus satisfies

$$\chi^2(\mathrm{P}_{h_\theta}, \mathrm{P}_{h_0}) = \int\left(\frac{\mathrm{P}_{h_\theta}}{\mathrm{P}_{h_0}}\right)d\mathrm{P}_{h_0} - 1 = \prod_{j=1}^{J}\frac{\exp(-n\lambda_j^2\gamma_j^2) - \exp(n\lambda_j^2\gamma_j^2)}{2} - 1,$$

where we define $\gamma_j = n^{-1/2}\tau_j^2\big(\sum_{j=1}^{J}\tau_j^4\big)^{-1/4}$. By Tsybakov [2009, Section 2.7.5] there exists a constant $c_1 > 0$ such that $\exp(-n\lambda_j^2\gamma_j^2) - \exp(n\lambda_j^2\gamma_j^2) \leq 2\exp\left(c_1 n^2\lambda_j^4\gamma_j^4\right)$. By making use of condition $\lambda_j \leq c\tau_j^{-1}$ for all $j \geq 1$ we obtain for some $c_2 > 0$ (which can be arbitrary small as $\tau_j \to \infty$) that

$$\chi^2(\mathrm{P}_{h_\theta}, \mathrm{P}_{h_0}) \leq \exp\Big(c_1 c n^2\sum_{j=1}^{J}\tau_j^4\gamma_j^4\Big) - 1 \leq \exp(c_2) - 1,$$

by the definition of γ_j. Consequently, the results follows by making use of inequality (A.1).

Finally, we derive specific rate results under the different measures of ill-posedness. Consider the mildly ill-posed case with the choice of $J \sim n^{2d_x/(4(p+\zeta)+d_x)}$ then J satisfies constraint (A.2) within a constant and

$$n^{-1}\Big(\sum_{j=1}^{J}\tau_j^4\Big)^{1/2}\sim n^{-1}\Big(\sum_{j=1}^{J}j^{4\zeta/d_x}\Big)^{1/2}\sim n^{-4p/(4(p+\zeta)+d_x)}.$$

In the severely ill-posed case, the choice of $J\sim(\log n-\zeta^{-1}(2p+1)\log\log n)^{1/\zeta}$ also satisfies (A.2) within a constant and

$$n^{-1}\Big(\sum_{j=1}^{J}\tau_j^4\Big)^{1/2}\sim n^{-1}\Big(\sum_{j=1}^{J}\exp\big(2j^{\zeta/d_x}\big)\Big)^{1/2}\sim(\log n)^{-2p/\zeta},$$

which completes the proof. □

B. Proofs of Adaptive ST Results in Subsection 3.3

We require some additional notation. We set $Z_i=(Y_i,X_i,W_i)$ and introduce a function

$$\begin{aligned}R(Z_i,Z_{i'},D_i)&=(Y_i-h_0(X_i))\mathbb{1}_{D_i}\,b^K(W_i)'A'GAb^K(W_{i'})(Y_{i'}-h_0(X_{i'}))\mathbb{1}_{D_{i'}}\\&\quad-\mathrm{E}_h[(Y-h_0(X))\mathbb{1}_D\,b^K(W)]'A'GA\,\mathrm{E}_h[b^K(W)(Y-h_0(X))\mathbb{1}_D]\end{aligned}$$

for any set D_i. We define $R_1(Z_i,Z_{i'}):=R(Z_i,Z_{i'},M_i)$ and $R_2(Z_i,Z_{i'}):=R(Z_i,Z_{i'},M_i^c)$ where $M_i=\{|Y_i-h_0(X_i)|\le M_n\}$ and $(M_n)_{n\ge1}$ is an increasing sequence satisfying $M_n=o\big(\sqrt{n}\,\overline{\zeta}^{-1}(\log\log\overline{J})^{-3/4}\big)$. Based on kernels R_l, where $l=1,2$, we introduce the U-statistic

$$\mathcal{U}_{J,l}=\frac{2}{n(n-1)}\sum_{1\le i<i'\le n}R_l(Z_i,Z_{i'}).$$

We make also use of the notation

$$\begin{aligned}\Lambda_1&=\Big(\frac{n(n-1)}{2}\,\mathrm{E}[R_1^2(Z_1,Z_2)]\Big)^{1/2}, &\Lambda_2&=n\sup_{\|\nu\|_Z\le1,\|\kappa\|_Z\le1}\mathrm{E}[R_1(Z_1,Z_2)\nu(Z_1)\kappa(Z_2)],\\\Lambda_3&=\big(n\sup_z|\,\mathrm{E}[R_1^2(Z_1,z)]|\big)^{1/2}, &\text{and}\quad\Lambda_4&=\sup_{z_1,z_2}|R_1(z_1,z_2)|.\end{aligned}$$

In addition to the data driven index set $\widehat{\mathcal{I}}_n$ given in (2.4) we now consider here a population analog

$$\mathcal{I}_n=\big\{J\le\overline{J}:\quad J=\underline{J}2^j\ \text{where}\ j=0,1,\dots,j_{\max}\big\}\tag{B.1}$$

where $\overline{J}$ satisfies the rate conditions captured in Assumption 5. From Chen and Christensen [2015, Theorem 3.2] we infer that $\widehat{J}_{\max}\le\overline{J}$ holds with probability approaching one and hence, we may restrict the index set to $\mathcal{I}_n$ in the following proofs. In particular, the empirical normalization term $\widehat{\eta}_J(\alpha)$ given in (2.5) becomes deterministic and we introduce the notation

$$\eta_J = \frac{q\big(\alpha/\#(\mathcal{I}_n), J\big) - J}{\sqrt{2J}}.$$

Proof of Theorem 3.3. For the proof of the result, we proceed in three steps. First, we bound the first type error of the test statistic

$$\widetilde{\mathtt{ST}}_n = \mathbb{1}\left\{\max_{J\in\mathcal{I}_n} |n\widehat{D}_J/(\eta'_J \mathrm{v}_J)| > 1\right\}$$

for some $\eta'_J > 0$. Second, we bound the second type error of $\widetilde{\mathtt{ST}}_n$ where η'_J is replaced by $\eta''_J > 0$. Let η'_J and η''_J be such that

$$\sqrt{2}\eta_J = \frac{\eta'_J}{1-c_0} = \frac{\eta''_J}{1+c_0}$$

for some constant $0 < c_0 < 1$. Finally, we show that the derived bounds in the previous steps are sufficient to control the first and second type error of our test statistic $\mathtt{ST}_n$.

Step 1: To control the first type error of the test statistic $\widetilde{\mathtt{ST}}_n$, we make use of the decomposition for all $h \in \mathcal{H}_0$:

$$\begin{aligned}
\mathrm{P}_h\left(\widetilde{\mathtt{ST}}_n = 1\right) &\le \mathrm{P}_h\left(\max_{J\in\mathcal{I}_n}\left\{|n\widehat{D}_J|/(\eta'_J\mathrm{v}_J)\right\} > 1\right). \\
&\le \mathrm{P}_h\Big(\max_{J\in\mathcal{I}_n}\Big|\frac{1}{\eta'_J\mathrm{v}_J(n-1)}\sum_{j=1}^J\sum_{i\neq i'} V_{ij}V_{i'j}\Big| \\
&\quad + \max_{J\in\mathcal{I}_n}\Big|\frac{1}{\eta'_J\mathrm{v}_J(n-1)}\sum_{i\neq i'} U_iU_{i'}b^K(W_i)'\Big(A'GA - \widehat{A}'G\widehat{A}\Big)b^K(W_{i'})\Big| > 1\Big) \\
&\le \underbrace{\mathrm{P}_h\Big(\max_{J\in\mathcal{I}_n}\big|n\mathcal{U}_{J,1}/(\eta'_J\mathrm{v}_J)\big| > \frac{1}{4}\Big)}_{=I} + \underbrace{\mathrm{P}_h\left(\max_{J\in\mathcal{I}_n}\big|n\mathcal{U}_{J,2}/(\eta'_J\mathrm{v}_J)\big| > \frac{1}{4}\right)}_{=II} \\
&\quad + \underbrace{\mathrm{P}_h\Big(\max_{J\in\mathcal{I}_n}\Big|\frac{1}{\eta'_J\mathrm{v}_J(n-1)}\sum_{i\neq i'} U_iU_{i'}b^K(W_i)'\big(A'GA - \widehat{A}'G\widehat{A}\big)b^K(W_{i'})\Big| > \frac{1}{2}\Big)}_{=III}
\end{aligned}$$

using that $U_i = Y_i - h_0(X_i)$ under $\mathcal{H}_0$. Consider I. From Lemma G.1 and Lemma G.2 we conclude for $u = 2\log\log J$ and $M_n = o\big(\sqrt{n}\,\overline{\zeta}^{-1}(\log\log\overline{J})^{-3/4}\big)$ that for all $J \in \mathcal{I}_n$ we have

$$\begin{aligned}
&\Lambda_1\sqrt{u} + \Lambda_2 u + \Lambda_3 u^{3/2} + \Lambda_4 u^2 \\
&\le \Lambda_1\sqrt{2\log\log J} + 2\Lambda_2\log\log J + \Lambda_3(2\log\log J)^{3/2} + 4\Lambda_4(\log\log J)^2 \\
&\le n\mathrm{v}_J\sqrt{\log\log J} + 4\overline{\sigma}^2 n s_J^{-2}\log\log J + \overline{\sigma}^2 n s_J^{-2}\,(2\log\log J)^{3/4} + 4ns_J^{-2}\,\sqrt{\log\log J}
\end{aligned}$$

for n sufficiently large. Lemma F.2 implies $s_J^{-2} \le \underline{\sigma}^{-2}\mathrm{v}_J$. Assumption 5 (ii) yields $\log\log J \le \underline{\sigma}^2\eta_J'/(16\overline{\sigma}^2)$ for all $J \in \mathcal{I}_n$ and n sufficiently large and hence,we obtain

$$\Lambda_1\sqrt{u} + \Lambda_2 u + \Lambda_3 u^{3/2} + \Lambda_4 u^2 \le \frac{n-1}{2}\mathrm{v}_J\eta_J'.$$

Consequently, the exponential inequality for degenerate U-statistics in Lemma G.1 with $u = 2\log\log J$ together with the definition of $\mathcal{I}_n$, i.e., $J = \underline{J}2^j$ for all $J \in \mathcal{I}_n$, yields

$$\begin{aligned}
I &\le \sum_{J\in\mathcal{I}_n} \mathrm{P}_h\left(\left|n\mathcal{U}_{J,1}\right| > \frac{\eta_J'}{4}\mathrm{v}_J\right) = \sum_{J\in\mathcal{I}_n}\mathrm{P}_h\left(\Big|\sum_{i<i'} R_1(Z_i, Z_{i'})\Big| \ge \frac{\eta_J'}{4}\frac{n-1}{2}\mathrm{v}_J\right) \\
&\le 6\sum_{J\in\mathcal{I}_n}\exp(-2\log\log J) \\
&= 6\sum_{J\in\mathcal{I}_n}(\log J)^{-2} \\
&\le 6\sum_{j\ge 0}(\log\underline{J} + j\log 2)^{-2} \\
&\le 6\Big((\log 2)(\log\underline{J} - \log 2)\Big)^{-1} = o(1)
\end{aligned}$$

by making use of the definition $\underline{J} = \lfloor 3\sqrt{\log\log n}\rfloor$. Consider II. By Markov's inequality we obtain

$$\begin{aligned}
II &\le \mathrm{E}_h\max_{J\in\mathcal{I}_n}\left|\frac{4}{\eta_J'\mathrm{v}_J(n-1)}\sum_{i<i'}U_i\mathbb{1}_{M_i^c}U_{i'}\mathbb{1}_{M_{i'}^c}b^K(W_i)'A'G_\mu Ab^K(W_{i'})\right| \\
&\le 4n\,\mathrm{E}_h\left|U\mathbb{1}_{\{|U|>M_n\}}\right|\,\mathrm{E}_h\left|U\mathbb{1}_{\{|U|>M_n\}}\right|\max_{J\in\mathcal{I}_n}\frac{\zeta_J^2\big\|G_\mu^{1/2}AG_b^{1/2}\big\|^2}{\eta_J'\mathrm{v}_J} \\
&\le 4nM_n^{-6}\big(\mathrm{E}_h[U^4]\big)^2\,\overline{\zeta}^2\max_{J\in\mathcal{I}_n}\frac{s_J^{-2}}{\eta_J'\mathrm{v}_J}
\end{aligned}$$

where the fourth moment of $U = Y - h_0(X)$ is bounded under Assumption 1 (ii). From Lemma F.2 we deduce $s_J^{-2} \le \underline{\sigma}^{-2}\mathrm{v}_J$. Thus, using definition $M_n = o\big(\sqrt{n}\,\overline{\zeta}^{-1}(\log\log\overline{J})^{-3/4}\big)$ gives

$$II = o\left(n^{-2}(\log\log\overline{J})^{9/2}\,\overline{\zeta}^8\right) = o(1)$$

where the last equation is due to rate restriction on $\overline{J}$ imposed Assumption 5 (i). Consider III. Lemma F.5 implies uniformly in $J \in \mathcal{I}_n$ that

$$\frac{1}{n-1}\sum_{i\neq i'} U_i U_{i'} b^K(W_i)'\big(A'GA - \widehat{A}'G\widehat{A}\big)b^K(W_{i'}) = o_p\left(\mathrm{v}_J \eta'_J\right)$$

and hence $III = o(1)$.

Step 2: We control the second type error of the test statistic $\widetilde{\mathsf{ST}}_n$ where η'_J is replaced by $\eta''_J > 0$. We may assume that there exists $\widetilde{J}$ with $\underline{J} \leq \widetilde{J} \leq n^{1/3}$ such that $\widetilde{J}^{-2p/d} \leq n^{-1}\mathrm{v}_{\widetilde{J}}$. Thus, by the construction of the set $\mathcal{I}_n$ we have that there exists $J^* \in \mathcal{I}_n$ such that $\widetilde{J} \leq J^* < 2\widetilde{J}$. We denote $K^* = K(J^*)$. We further evaluate for all $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ that

$$\mathrm{P}_h\left(\widetilde{\mathsf{ST}}_n = 0\right) = \mathrm{P}_h\left(n\,\widehat{D}_J \leq \eta''_J\,\mathrm{v}_J \text{ for all } J \in \mathcal{I}_n\right) \leq \mathrm{P}_h\left(n\,\widehat{D}_{J^*} \leq \eta''_{J^*}\,\mathrm{v}_{J^*}\right).$$

We make use of the notation $B_J = \|\,\mathrm{E}_h[V^J]\|_2^2 - \|h - h_0\|_\mu^2$ and obtain uniformly over $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ that

$$\begin{aligned}
&\mathrm{P}_h\left(n\,\widehat{D}_{J^*} \leq \eta''_{J^*}\,\mathrm{v}_{J^*}\right) = \mathrm{P}_h\left(\|\,\mathrm{E}_h[V^{J^*}]\|_2^2 - \widehat{D}_{J^*} > \|\,\mathrm{E}_h[V^{J^*}]\|_2^2 - \frac{\eta''_{J^*}\,\mathrm{v}_{J^*}}{n}\right)\\
&\leq \underbrace{\mathrm{P}_h\left(\Big|\frac{8}{n^2}\sum_{j=1}^{J^*}\sum_{i<i'}\big(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\big)\Big| > \|h-h_0\|_\mu^2 - \frac{\eta''_{J^*}\,\mathrm{v}_{J^*}}{n} + B_{J^*}\right)}_{IV}\\
&+ \underbrace{\mathrm{P}_h\left(\Big|\frac{8}{n^2}\sum_{i<i'}(Y_i - h_0(X_i))(Y_{i'} - h_0(X_{i'})b^{K^*}(W_i)'\Big(A'GA - \widehat{A}'G\widehat{A}\Big)b^{K^*}(W_{i'})\Big| > \|h-h_0\|_\mu^2 - \frac{\eta''_{J^*}\,\mathrm{v}_{J^*}}{n} + B_{J^*}\right)}_{V}
\end{aligned}$$

using $n/(n-1) \leq 2$ for all $n \geq 2$. Consider IV. We first derive an upper bound for the term B_{J^*}. The definitions of V^J and Q_J imply

$$\|\,\mathrm{E}_h[V^{J^*}]\|^2 = \mathrm{E}_h[(Y - h_0(X))b^{K^*}(W)']A'GA\,\mathrm{E}_h[(Y - h_0(X))b^{K^*}(W)] = \|Q_{J^*}(h - h_0)\|_\mu^2.$$

Consequently, from Lemma F.3 we infer

$$|B_{J^*}| = \big|\|Q_{J^*}(h - h_0)\|_\mu^2 - \|h - h_0\|_\mu^2\big| \leq C_B\,\mathsf{r}_n^2$$

for some constant C_B, due to the definition of J^*. To establish an upper bound of IV, we make use of inequality (F.3) together with Markov's inequality which yields

$$IV = O\left(\frac{n^{-1}\big\|\langle Q_{J^*}(h-h_0), \psi^{J^*}\rangle'_\mu (G_b^{-1/2}S)^-_l\big\|^2 + n^{-2}\mathrm{v}_{J^*}^2}{\big(\|h-h_0\|_\mu^2 - \eta''_{J^*}\,n^{-1}\mathrm{v}_{J^*} + B_{J^*}\big)^2}\right). \tag{B.2}$$

In the following, we distinguish between two cases. First, consider the case where $n^{-2}\mathrm{v}_{J^*}^2$

dominates the summand in the numerator. Assumption 5 (ii) implies $\eta''_{J^*} \leq C_\eta \sqrt{\log\log n}$ for some constant $C_\eta > 0$. For any $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ we have $\|h - h_0\|_\mu^2 \geq \delta^\circ\, \mathsf{r}_n^2$ and hence, we obtain the lower bound

$$\|h - h_0\|_\mu^2 - \eta''_{J^*}\, n^{-1} \mathrm{v}_{J^*} + B_{J^*} \geq (\delta^\circ - C_\eta - C_B)\, \mathsf{r}_n^2 \geq c_0 \mathsf{r}_n^2$$

for some constant $c_0 := \delta^\circ - C_\eta - C_B$ which is positive for any $\delta^\circ > C_\eta + C_B$. From inequality (B.2) we infer

$$IV \leq O\left(\frac{\mathrm{v}_{J^*}^2}{c_0^2\, \mathsf{r}_n^4\, n^2}\right) = o(1).$$

Second, consider the case where $n^{-1}\big\|\langle Q_{J^*}(h - h_0), \psi^{J^*}\rangle'_\mu (G_b^{-1/2} S)_l^-\big\|^2$ dominates. Now using $\|(G_b^{-1/2} S G^{-1/2})_l^-\| = s_{J^*}^{-1}$ together with Assumption 2 (iii), i.e., the eigenvalues of G are uniformly bounded away from zero, we obtain

$$\begin{aligned} n^{-1}\big\|\langle Q_{J^*}(h - h_0), \psi^{J^*}\rangle'_\mu (G_b^{-1/2} S)_l^-\big\|^2 &= n^{-1}\big\|\langle Q_{J^*}(h - h_0), \psi^{J^*}\rangle'_\mu G^{-1/2} (G_b^{-1/2} S G^{-1/2})_l^-\big\|^2 \\ &= O\Big(n^{-1} s_{J^*}^{-2} \big\|\langle Q_{J^*}(h - h_0), \psi^{J^*}\rangle_\mu\big\|^2\Big) \\ &= O\Big(n^{-1} s_{J^*}^{-2} \big(\|h - h_0\|_\mu^2 + (J^*)^{-2p/d}\big)\Big), \end{aligned}$$

where the last bound is due to Lemma F.3. For any $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ we have $\|h - h_0\|_\mu^2 \geq \delta^\circ\, \mathsf{r}_n^2 > \delta^\circ n^{-1} \mathrm{v}_{J^*}$ and hence, obtain the lower bound

$$\|h - h_0\|_\mu^2 - \eta''_{J^*}\, n^{-1} \mathrm{v}_{J^*} + B_{J^*} \geq \left(1 - \frac{1}{\delta^\circ} - \frac{C_B}{\delta^\circ}\right) \|h - h_0\|_\mu^2 \geq c_1 \|h - h_0\|_\mu^2$$

for some constant $c_1 = 1 - (1 + C_B)/\delta^\circ$ which is positive for any $\delta^\circ > 1 + C_B$. Hence, inequality (B.2) yields for all $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ that

$$IV = O\Big(n^{-1} s_{J^*}^{-2} \Big(\frac{1}{\|h - h_0\|_\mu^2} + \frac{1}{\|h - h_0\|_\mu^4 (J^*)^{2p/d}}\Big)\Big) = O\Big(n^{-1} s_{J^*}^{-2} \mathsf{r}_n^{-2}\Big) = o(1).$$

Finally, $V = o(1)$ by making use of Lemma F.4.

Step 3: Finally, we account for estimation of the normalization factor v_J. Now since $|\widehat{\mathrm{v}}_J - \mathrm{v}_J| \leq c_0 \mathrm{v}_J$ wpa1 for some $0 < c_0 < 1$ for all $J \in \mathcal{I}_n$ by Lemma F.6 it holds $(1 - c_0)\mathrm{v}_J \leq \widehat{\mathrm{v}}_J \leq (1 + c_0)\mathrm{v}_J$ wpa1. Hence, we control the first type error of the test $\mathtt{ST}_n$

as follows, using for any $\widetilde{J} \in \mathcal{I}_n$ and for all $h \in \mathcal{H}_0$ that

$$
\begin{aligned}
\mathrm{P}_h\left(\mathtt{ST}_n = 1\right) &\leq \mathrm{P}_h\left(\mathtt{ST}_n = 1, \quad \widehat{\mathrm{v}}_J \geq (1-c_0)\mathrm{v}_J \text{ for all } J \in \mathcal{I}_n\right) \\
&\quad + \mathrm{P}_h\left(\widehat{\mathrm{v}}_J < (1-c_0)\mathrm{v}_J \text{ for all } J \in \mathcal{I}_n\right) \\
&\leq \mathrm{P}_h\left(\max_{J\in\mathcal{I}_n}\left\{|\widehat{D}_J|/(\sqrt{2}\eta_J\widehat{\mathrm{v}}_J)\right\} > 1, \quad \widehat{\mathrm{v}}_J \geq (1-c_0)\mathrm{v}_J \text{ for all } J \in \mathcal{I}_n\right) \\
&\quad + \mathrm{P}_h\left(\widehat{\mathrm{v}}_{\widetilde{J}} < (1-c_0)\mathrm{v}_{\widetilde{J}}\right) \\
&\leq \mathrm{P}_h\left(\max_{J\in\mathcal{I}_n}\left\{|\widehat{D}_J|/(\sqrt{2}\eta_J\mathrm{v}_J)\right\} > 1-c_0\right) + o(1) \\
&= \mathrm{P}_h\left(\max_{J\in\mathcal{I}_n}\left\{|\widehat{D}_J|/(\eta_J'\mathrm{v}_J)\right\} > 1\right) + o(1) = o(1)
\end{aligned}
$$

where the last equation is due to Step 1. To bound the second error term of the test $\mathtt{ST}_n$ recall the definition of $J^* \in \mathcal{I}_n$ given in Step 2. We evaluate for all $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ that

$$
\begin{aligned}
\mathrm{P}_h\left(\mathtt{ST}_n = 0\right) &\leq \mathrm{P}_h\left(|\widehat{D}_{J^*}| \leq \sqrt{2}\eta_{J^*}\,\widehat{\mathrm{v}}_{J^*}\right) \\
&\leq \mathrm{P}_h\left(|\widehat{D}_{J^*}| \leq \sqrt{2}\eta_{J^*}\,\widehat{\mathrm{v}}_{J^*}, \quad \widehat{\mathrm{v}}_{J^*} \leq (1+c_0)\mathrm{v}_{J^*}\right) + \mathrm{P}_h\left(\widehat{\mathrm{v}}_{J^*} > (1+c_0)\mathrm{v}_{J^*}\right) \\
&= \mathrm{P}_h\left(|\widehat{D}_{J^*}| \leq \sqrt{2}\eta_{J^*}\,(1+c_0)\mathrm{v}_{J^*}\right) + o(1) \\
&= \mathrm{P}_h\left(|\widehat{D}_{J^*}| \leq \eta_{J^*}''\,\mathrm{v}_{J^*}\right) + o(1) = o(1),
\end{aligned}
$$

where the last equation is due to Step 2. □

Proof of Corollary 3.2. Theorem 3.3 establishes the rate $\mathsf{r}_n^2 = n^{-1}\sqrt{\log\log n}\, s_{J_0}^{-2}\sqrt{J_0}$ where J_0 satisfies $J_0^{-2p/d_x} \sim n^{-1}\sqrt{\log\log n}\, s_{J_0}^{-2}\sqrt{J_0}$. Consequently, following the proof of Corollary 3.1, the result immediately follows. □

Supplement to "Adaptive, Rate-Optimal Testing in Instrumental Variables Models"

Christoph Breunig Xiaohong Chen

First version: August 2018, Revised June 16, 2020

This supplementary appendix contains materials to support our paper. Appendix C contains robustness checks using bootstrap critical values for the empirical illustrations. Appendix D provides proofs of our adaptive ST results under composite null hypothesis. Appendix E establishes the results for our adaptive IT for conditional moment restrictions. Appendix F provides an upper bound for quadratic distance estimation which is essential for our minimax testing results. It also contains further technical Lemmas. Finally, Appendix G gathers an exponential inequality for U-statistics.

C. Empirical Results based on Bootstrap Critical Values

In the section, we present robustness checks of our empirical findings when the critical values of our adaptive test are computed using bootstrap. For the bootstrap critical values $\widehat{\eta}_J(\alpha)$, the i.i.d. bootstrap weight is generated according to $\omega \sim \mathcal{N}(0,1)$, independently from the original data sample, according to Remark 3.2 for constrained inequality tests. We use 1000 bootstrap iterations. We see below that all the empirical findings based on bootstrap critical values coincide with the empirical results based on our simple adaptive critical values reported in the main paper.

Income	H_0: h is decreasing in P_o				H_0: h is increasing in P_o			
groups	$\widehat{\mathcal{W}}_{\widehat{\mathcal{J}}}$	p val.	reject H_0?	$\widehat{\mathcal{J}}$	$\widehat{\mathcal{W}}_{\widehat{\mathcal{J}}}$	p val.	reject H_0?	$\widehat{\mathcal{J}}$
low	0.497	0.094	no	$\{8\}$	1.695	0.003	yes	6
middle	0.292	0.175	no	$\{10\}$	4.918	0.000	yes	$\{6,8,10\}$
high	1.471	0.002	yes	$\{8,10,12\}$	6.070	0.000	yes	$\{6,8,10,12\}$

Table A: Adaptive bootstrap testing for monotonic demand for organic products.

Table A reports bootstrap adaptive testing results for monotonic multi-product demand in Subsection 5.1.1. It replicates Table 2.

Income	H_0: h is decreasing in P				H_0: h is increasing in P			
groups	$\widehat{\mathcal{W}}_{\hat{\mathcal{J}}}$	p val.	reject H_0?	$\hat{\mathcal{J}}$	$\widehat{\mathcal{W}}_{\hat{\mathcal{J}}}$	p val.	reject H_0?	$\hat{\mathcal{J}}$
low	0.408	0.159	no	$\{8\}$	2.611	0.000	yes	$\{8\}$
high	15.256	0.000	yes	$\{8, 10\}$	11.894	0.000	yes	$\{10\}$

Table B: Adaptive bootstrap testing for monotonic gasoline demand.

Table B reports bootstrap adaptive testing results for monotonic gasoline demand in Subsection 5.1.2. It replicates Table 3.

	H_0: h is increasing				H_0: h is decreasing			
Goods	$\widehat{\mathcal{W}}_{\hat{\mathcal{J}}}$	p value	reject H_0?	$\hat{\mathcal{J}}$	$\widehat{\mathcal{W}}_{\hat{\mathcal{J}}}$	p value	reject H_0?	$\hat{\mathcal{J}}$
"food in"	1.449	0.003	yes	$\{3\}$	-0.262	0.947	no	$\{6\}$
"fuel"	3.936	0.000	yes	$\{3\}$	0.327	0.677	no	$\{3\}$
"travel"	2.079	0.001	yes	$\{3\}$	0.246	0.144	no	$\{3\}$
"leisure"	0.241	0.152	no	$\{6\}$	2.851	0.000	yes	$\{3, 4\}$

Table C: Adaptive bootstrap testing for monotonicity of Engel curves.

Table C reports bootstrap adaptive testing results for monotonic Engel curves in Subsection 5.1.3. It replicates Table 4.

	H_0: h is linear				H_0: h is quadratic			
Goods	$\widehat{\mathcal{W}}_{\hat{\mathcal{J}}}$	p value	reject H_0?	$\hat{\mathcal{J}}$	$\widehat{\mathcal{W}}_{\hat{\mathcal{J}}}$	p value	reject H_0?	$\hat{\mathcal{J}}$
"food in"	-0.177	0.785	no	$\{3\}$	0.240	0.159	no	$\{3\}$
"fuel"	1.141	0 .009	yes	$\{3\}$	-0.078	0.502	no	$\{4\}$
"travel"	1.260	0.005	yes	$\{3\}$	-0.098	0.539	no	$\{5\}$
"leisure"	0.482	0.077	no	$\{6\}$	0.408	0.092	no	$\{3\}$

Table D: Adaptive bootstrap testing for linear/quadratic Engel curves.

Table D reports bootstrap adaptive testing results for either linear or quadratic Engel curves in Subsection 5.1.3. It replicates Table 5.

D. Proofs of Adaptive ST Results Under Composite Null Hypothesis in Subsection 3.4

Recall the definition of the deterministic index set $\mathcal{I}_n$ in (B.1) satisfying $\widehat{\mathcal{I}}_n \subset \mathcal{I}_n$ with probability approaching one. Also recall the notation $\widetilde{b}^K(\cdot) = G_b^{-1/2} b^K(\cdot)$. Below, we denote by $C > 0$ a generic constant that may be different in different uses. For simplicity of notation, we write v_J instead of $v_J(h^{\mathrm{R}})$ and $\widehat{\eta}_J(\alpha)$ instead of $\widehat{\eta}_J$. Below, η_J denotes a deterministic sequence satisfying $\eta_{\overline{J}} = O(\sqrt{\log\log n})$ and $(\log\log J)^c \leq \eta_J$ for some constant $c > 1$ and for all $\underline{J} \leq J \leq \overline{J}$.

Proof of Theorem 3.4. For the proof of the result, we proceed in three steps. First, we bound the first type error of the test statistic

$$\widetilde{\mathtt{ST}}_n = \mathbb{1}\left\{ \max_{J \in \mathcal{I}_n} |n\widehat{D}_J(\widehat{h}_J^{\mathrm{R}})/(\eta_J \widehat{\mathrm{v}}_J(\widehat{h}_J^{\mathrm{R}}))| > \sqrt{2} \right\}.$$

Second, we bound the second type error of $\widetilde{\mathtt{ST}}_n$. Third, we show that these steps are sufficient to control the first and second type error of our test statistic $\widehat{\mathtt{ST}}_n$.

Step 1: We control the first type error of the test statistic $\widetilde{\mathtt{ST}}_n$ by making use of the decomposition for all $h \in \mathcal{H}_0$:

$$\begin{aligned} \mathrm{P}_h\Big(\widetilde{\mathtt{ST}}_n = 1\Big) \leq & \underbrace{\mathrm{P}_h\Big(\max_{J \in \mathcal{I}_n} \frac{n\,\widehat{D}_J(h^{\mathrm{R}})}{\eta_J\,\widehat{\mathrm{v}}_J(h^{\mathrm{R}})} > \sqrt{1/2}\Big)}_{I} \\ & + \underbrace{\mathrm{P}_h\Big(n \max_{J \in \mathcal{I}_n} \sup_{h \in \mathcal{H}_J^{\mathrm{R}}} \big|\widehat{D}_J(h) - \widehat{D}_J(h^{\mathrm{R}})\big| / (\eta_J\,\widehat{\mathrm{v}}_J(h^{\mathrm{R}})) > \sqrt{1/8}\Big)}_{II} \\ & + \underbrace{\mathrm{P}_h\Big(\max_{J \in \mathcal{I}_n} \Big|\frac{n\,\widehat{D}_J(\widehat{h}_J)}{\eta_J\,\widehat{\mathrm{v}}_J(h^{\mathrm{R}})}\Big| \max_{J \in \mathcal{I}_n} \Big|1 - \frac{\widehat{\mathrm{v}}_J(h^{\mathrm{R}})}{\widehat{\mathrm{v}}_J(\widehat{h}_J)}\Big| > \sqrt{1/8}\Big)}_{III}. \end{aligned}$$

We have $I = o(1)$ due to the proof of Theorem 3.1. Consider II. We may assume that $\widehat{h}_J^{\mathrm{R}} \in \mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})$ due to Assumption 6 (i). The definition of the estimator $\widehat{D}_J(h)$ implies for all $J \in \mathcal{I}_n$ and $h \in \mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})$ that

$$\begin{aligned}
&\frac{n}{\eta_J \mathrm{v}_J}\big(\widehat{D}_J(h) - \widehat{D}_J(h^{\mathrm{R}})\big) \\
&\quad= \frac{1}{\eta_J \mathrm{v}_J (n-1)} \sum_{i \neq i'} (h - h^{\mathrm{R}})(X_i)(h - h^{\mathrm{R}})(X_{i'}) b^K(W_i)' A' G A b^K(W_{i'}) \\
&\qquad+ \frac{1}{\eta_J \mathrm{v}_J (n-1)} \sum_{i \neq i'} (h - h^{\mathrm{R}})(X_i)(h - h^{\mathrm{R}})(X_{i'}) b^K(W_i)' \Big(A' G A - \widehat{A}' G \widehat{A}\Big) b^K(W_{i'}) \\
&\qquad+ \frac{2}{\eta_J \mathrm{v}_J (n-1)} \sum_{i \neq i'} \big(Y_i - h^{\mathrm{R}}(X_i)\big)(h - h^{\mathrm{R}})(X_{i'}) b^K(W_i)' A' G A b^K(W_{i'}) \\
&\qquad+ \frac{2}{\eta_J \mathrm{v}_J (n-1)} \sum_{i \neq i'} \big(Y_i - h^{\mathrm{R}}(X_i)\big)(h - h^{\mathrm{R}})(X_{i'}) b^K(W_i)' \Big(A' G A - \widehat{A}' G \widehat{A}\Big) b^K(W_{i'}) \\
&\qquad= T_{1,J}(h) + T_{2,J}(h) + T_{3,J}(h) + T_{4,J}(h).
\end{aligned}$$

Consider $T_{1,J}(h)$. We obtain

$$\begin{aligned}
|T_{1,J}(h)| \leq &\frac{1}{\eta_J \mathrm{v}_J} \Big\| n^{-1/2} \sum_{i=1}^{n} (h - h^{\mathrm{R}})(X_i) b^K(W_i)' A' G^{1/2} \Big\|^2 \\
&+ \frac{1}{n \eta_J \mathrm{v}_J} \sum_{i=1}^{n} \Big\| (h - h^{\mathrm{R}})(X_i) b^K(W_i)' A' G^{1/2} \Big\|^2 \\
= &\, T_{11,J}(h) + T_{12,J}(h).
\end{aligned}$$

Consider $T_{11,J}(h)$. We obtain

$$\begin{aligned}
&\mathrm{P}_h\Big(\max_{J \in \mathcal{I}_n} \sup_{h \in \mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})} T_{11,J}(h) > 1\Big) \\
&\leq \mathrm{P}_h\Big(\exists J \in \mathcal{I}_n : \sup_{h \in \mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})} \Big\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big((h - h^{\mathrm{R}})(X_i) b^K(W_i)' - \mathrm{E}\big[(h - h^{\mathrm{R}})(X) b^K(W)'\big]\Big) A' G^{1/2} \Big\|^2 \\
&\qquad\qquad > \frac{\eta_J \mathrm{v}_J}{2} - \sup_{h \in \mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})} \big\| G^{1/2} A\, \mathrm{E}\big[(h - h^{\mathrm{R}})(X) b^K(W)\big] \big\|^2\Big) \\
&\leq \mathrm{P}_h\Big(\exists J \in \mathcal{I}_n : \sup_{h \in \mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})} \Big\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big((h - h^{\mathrm{R}})(X_i) b^K(W_i)' - \mathrm{E}\big[(h - h^{\mathrm{R}})(X) b^K(W)'\big]\Big) A' G^{1/2} \Big\|^2 \\
&\qquad\qquad > (1 - c_0) \eta_J \mathrm{v}_J\Big),
\end{aligned}$$

where the second inequality is due Lemma F.3, i.e., that uniformly in $h \in \mathcal{H}_J^{\mathrm{R}}$:

$$\big\| G^{1/2} A\, \mathrm{E}\big[(h - h^{\mathrm{R}})(X) b^K(W)\big] \big\|^2 = \| Q_J (h - h^{\mathrm{R}}) \|_\mu^2 \leq C(\Delta_{J,n}^2 + J^{-2p/d})$$

and for n sufficiently large that $2C(\Delta_{J,n}^2 + J^{-2p/d}) \leq c_0 \eta_J \mathrm{v}_J$ for some $0 < c_0 < 1$ and all $J \in \mathcal{I}_n$. Let s_j^{-1}, $1 \leq j \leq J$, be the nondecreasing singular values of $G^{1/2} A G_b^{1/2}$. Further, let e_j be the unit vector with 1 at the j–th position. We bound for all $h \in \mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})$ and all $j \geq 1$

$$\begin{aligned}
\mathrm{E} \sup_{h \in \mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})} & \left|(h - h^{\mathrm{R}})(X)\widetilde{b}^K(W)'\mathrm{diag}(s_1^{-1}, \ldots, s_J^{-1})e_j\right|^2 \\
& \leq \sup_{h \in \mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})} \|h - h^{\mathrm{R}}\|_\infty^2 \, \mathrm{E}\left|\widetilde{b}^K(W)'\mathrm{diag}(s_1^{-1}, \ldots, s_J^{-1})e_j\right|^2 \\
& \leq \Delta_{J,n}^2 \left\|\mathrm{diag}(s_1^{-1}, \ldots, s_J^{-1})e_j\right\|^2 \\
& \leq s_j^{-2}\Delta_{J,n}^2
\end{aligned}$$

by the definition of $\mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})$. Consequently, Markov's inequality together with the proof of Chen and Pouzo [2012, Lemma C.1] yields

$$\begin{aligned}
& \mathrm{P}_h\Big(\max_{J \in \mathcal{I}_n} \sup_{h \in \mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})} T_{11,J}(h) > 1\Big) \leq \sum_{J \in \mathcal{I}_n} \mathrm{P}_h\Big(\sup_{h \in \mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})} T_{11,J}(h) > 1\Big) \\
& \leq \sum_{J \in \mathcal{I}_n} \frac{1}{(1 - c_0)\eta_J \mathrm{v}_J} \mathrm{E} \sup_{h \in \mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})} \left\|\frac{1}{\sqrt{n}} \sum_{i=1}^n \Big((h - h^{\mathrm{R}})(X_i)b^K(W_i)' - \mathrm{E}\left[(h - h^{\mathrm{R}})(X)b^K(W)'\right]\Big)A'G^{1/2}\right\|^2 \\
& \leq C(1 - c_0)^{-1} \sum_{J \in \mathcal{I}_n} \frac{1}{\eta_J \mathrm{v}_J}\Big(\sum_{j=1}^J s_j^{-2}\Big)\Delta_{J,n}^2 \left(\int_0^1 \sqrt{1 + \log N_{[]}(tC, \mathcal{H}_J^{\mathrm{R}}, \|\cdot\|_\mu)}dt\right)^2 \\
& \leq C\underline{\sigma}^{-2}(1 - c_0)^{-1} \sum_{J \in \mathcal{I}_n} \frac{\Delta_{J,n}^2}{\eta_J} C_{J,n}^2,
\end{aligned}$$

for n sufficiently large, using $\sum_{j=1}^J s_j^{-2} \leq 2\underline{\sigma}^{-2}\sqrt{\mathrm{v}_J}$ by Lemma F.2. Consequently, the rate condition imposed in Assumption 6 (i) implies

$$\mathrm{P}_h\Big(\max_{J \in \mathcal{I}_n} \sup_{h \in \mathcal{H}_J^{\mathrm{R}}(\Delta_{J,n})} T_{11,J}(h) > 1\Big) = o(1). \tag{D.1}$$

Consider $T_{12,J}(h)$. Using the notation $\widetilde{b}^K(\cdot) = G_b^{-1/2}b^K(\cdot)$ we obtain

$$\begin{aligned}
\max_{J \in \mathcal{I}_n} \sup_{h \in \mathcal{H}_J^{\mathrm{R}}} T_{12,J}(h) & \leq \max_{J \in \mathcal{I}_n} \Big(\sup_{h \in \mathcal{H}_J^{\mathrm{R}}} \|h - h^{\mathrm{R}}\|_\infty \sup_w \|\widetilde{b}^K(w)\| \big\|\big(G_b^{-1/2}SG^{-1/2}\big)_l^- \big\|\Big)^2 / (\eta_J \mathrm{v}_J) \\
& \leq \max_{J \in \mathcal{I}_n} \frac{\big(\Delta_{J,n}\zeta_J s_J^{-1}\big)^2}{\eta_J \mathrm{v}_J} = o(1).
\end{aligned}$$

Consider $T_{2,J}(h)$. For all $J \in \mathcal{I}_n$ and $h \in \mathcal{H}_J^{\mathrm{R}}$ we evaluate

$$\begin{aligned}
|T_{2,J}(h)| & \leq \frac{2n}{\eta_J \mathrm{v}_J}\left\|\frac{1}{n}\sum_{i=1}^n (h - h^{\mathrm{R}})(X_i)\widetilde{b}^K(W_i) - \mathrm{E}\left[(h - h^{\mathrm{R}})(X)\widetilde{b}^K(W)\right]\right\|^2 \left\|G^{1/2}(\widehat{A} - A)G_b^{1/2}\right\|^2 \\
& \quad + \frac{2n}{\eta_J \mathrm{v}_J}\left\|\mathrm{E}\left[(h - h^{\mathrm{R}})(X)\widetilde{b}^K(W)\right]G^{1/2}(\widehat{A} - A)G_b^{1/2}\right\|^2 \\
& = 2T_{21,J}(h) + 2T_{22,J}(h).
\end{aligned}$$

Consequently, using Chen and Christensen [2018, Lemma F.10 (b)] (with G_ψ replaced by G), i.e., $\big\|G^{1/2}(\widehat{A}-A)G_b^{1/2}\big\|^2 = O_p\big(n^{-1}s_J^{-4}\zeta_J^2(\log J)\big)$ we obtain

$$\begin{aligned}
\mathrm{P}_h\Big(\max_{J\in\mathcal{I}_n}\sup_{h\in\mathcal{H}_J^{\mathrm{R}}} T_{21,J}(h)>1\Big) &\leq \sum_{J\in\mathcal{I}_n}\mathrm{P}_h\Big(\sup_{h\in\mathcal{H}_J^{\mathrm{R}}} T_{21,J}(h)>1\Big)\\
&\leq C\sum_{J\in\mathcal{I}_n}\bigg(\int_0^1\sqrt{1+\log N_{[]}(wC,\mathcal{H}_J^{\mathrm{R}},\|\cdot\|_\mu)}dw\bigg)^2\frac{K(J)}{\eta_J \mathrm{v}_J}\Delta_{J,n}^2 n^{-1}s_J^{-4}\zeta_J^2(\log J)\\
&\leq Cn^{-1}s_{\overline{J}}^{-2}\overline{\zeta}^4(\log n)\sum_{J\in\mathcal{I}_n}C_{J,n}^2\frac{\Delta_{J,n}^2}{\eta_J},
\end{aligned}$$

again using by Lemma F.2. By Assumption 5 (i) we have $s_{\overline{J}}^{-1}\overline{\zeta}^2\sqrt{(\log n)/n}=o(1)$ and consequently,

$$\mathrm{P}_h\Big(\max_{J\in\mathcal{I}_n}\sup_{h\in\mathcal{H}_J^{\mathrm{R}}} T_{2,J}(h)>1\Big)=o\Big(\sum_{J\in\mathcal{I}_n}C_{J,n}^2\frac{\Delta_{J,n}^2}{\eta_J}\Big)=o(1),$$

where the last equation is due to Assumption 6 (i). Consider $T_{22,J}(h)$. We make use of the inequality

$$T_{22,J}(h)\leq n\|\Pi_K T(h-h^{\mathrm{R}})\|_{L^2(W)}^2\big\|G^{1/2}(\widehat{A}-A)G_b^{1/2}\big\|^2.$$

Now Assumption 3 implies $\|\Pi_K T(h-h^{\mathrm{R}})\|_{L^2(W)}=O(s_J\|h-h^{\mathrm{R}}\|_\mu)$ and thus, we obtain

$$\max_{J\in\mathcal{I}_n}\sup_{h\in\mathcal{H}_J^{\mathrm{R}}} T_{22,J}(h)=o\left(\max_{J\in\mathcal{I}_n}\Delta_{J,n}^2\zeta_J^2(\log J)\right)=o(1),$$

by Assumption 6 (i). The bound of $T_{3,J}(h)$ and $T_{4,J}(h)$ follow analogously. Consider III. We make use of the inequality

$$III\leq \mathrm{P}_h\Big(\max_{J\in\mathcal{I}_n}\Big|\frac{n\,\widehat{D}_J(\widehat{h}_J^{\mathrm{R}})}{\eta_J\,\widehat{\mathrm{v}}_J(h^{\mathrm{R}})}\Big|>\sqrt{1/32}\Big)+\mathrm{P}_h\Big(\max_{J\in\mathcal{I}_n}\Big|1-\frac{\widehat{\mathrm{v}}_J(h^{\mathrm{R}})}{\widehat{\mathrm{v}}_J(\widehat{h}_J^{\mathrm{R}})}\Big|>\sqrt{1/32}\Big)$$

where the first term on the right hand side tends to zero which follows immediately from the upper bounds derived for I and II. Consequently, from Lemma F.7 we infer $III=o(1)$.

Step 2: We control the second type error of the test statistic $\widetilde{\mathtt{ST}}_n$ where η_J' is replaced by $\eta_J''>0$. Let J^* be as in the proof of Theorem 3.1. We thenobtain uniformly over

$h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ that

$$\begin{aligned}
\mathrm{P}_h\left(\widetilde{\mathtt{ST}}_n = 0\right) &\leq \mathrm{P}_h\left(n\,\widehat{D}_{J^*}(\widehat{h}^{\mathrm{R}}_{J^*}) \leq \eta''_{J^*}\,\mathrm{v}_{J^*}\right)\\
&\leq \mathrm{P}_h\left(\big|\|\,\mathrm{E}_h[V^{J^*}]\|_2^2 - \widehat{D}_{J^*}(h^{\mathrm{R}})\big| > \|\,\mathrm{E}_h[V^{J^*}]\|_2^2/2 - \frac{\eta''_{J^*}\,\mathrm{v}_{J^*}}{2n}\right)\\
&\quad + \mathrm{P}_h\left(\big|\widehat{D}_{J^*}(\widehat{h}^{\mathrm{R}}_{J^*}) - \widehat{D}_{J^*}(h^{\mathrm{R}})\big| > \|\,\mathrm{E}_h[V^{J^*}]\|_2^2/2 - \frac{\eta''_{J^*}\,\mathrm{v}_{J^*}}{2n}\right)
\end{aligned}$$

where the first summand on the right hand side tends to zero following the proof of Theorem 3.1. The second summand tends to zero analogously to the previous step of this proof.

Step 3: Finally, we account for estimation of the normalization factor v_J. We control the first type error of the test $\widehat{\mathtt{ST}}_n$ as follows, using for any $\widetilde{J} \in \mathcal{I}_n$ and $h \in \mathcal{H}_0$ by making use of Assumption 6 (iii)

$$\begin{aligned}
\mathrm{P}_h\left(\widehat{\mathtt{ST}}_n = 1\right) &\leq \mathrm{P}_h\left(\max_{J\in\mathcal{I}_n}\left\{|\widehat{D}_J(\widehat{h}^{\mathrm{R}}_J)|/(\widehat{\eta}_J \mathrm{v}_J)\right\} > 1, \quad \widehat{\eta}_J \geq C\eta_J \text{ for all } J \in \mathcal{I}_n\right)\\
&\quad + \mathrm{P}_h\left(\widehat{\eta}_J < C\eta_J \text{ for all } J \in \mathcal{I}_n\right) + o(1)\\
&\leq \mathrm{P}_h\left(\max_{J\in\mathcal{I}_n}\left\{|\widehat{D}_J(\widehat{h}^{\mathrm{R}}_J)|/(C\eta_J \mathrm{v}_J)\right\} > 1\right) + \mathrm{P}_h\left(\widehat{\eta}_{\widetilde{J}} < C\eta_{\widetilde{J}}\right) + o(1)\\
&= o(1)
\end{aligned}$$

where the last equation is due to Step 1. To bound the second error term of the test $\mathtt{ST}_n$ recall the definition of $J^* \in \mathcal{I}_n$ introduced in Step 2. We evaluate for all $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ by making use of Assumption 6 (iii) that

$$\begin{aligned}
\mathrm{P}_h\left(\widehat{\mathtt{ST}}_n = 0\right) &\leq \mathrm{P}_h\left(|\widehat{D}_{J^*}(\widehat{h}^{\mathrm{R}}_J)| \leq \widehat{\eta}_{J^*}\,\mathrm{v}_{J^*}, \quad \widehat{\eta}_{J^*} \leq c\,\eta_{J^*}\right) + \mathrm{P}_h\left(\widehat{\eta}_{J^*} > c\,\eta_{J^*}\right)\\
&= \mathrm{P}_h\left(|\widehat{D}_{J^*}(\widehat{h}^{\mathrm{R}}_J)| \leq c\,\eta_{J^*}\,\mathrm{v}_{J^*}\right) + o(1)\\
&= o(1),
\end{aligned}$$

where the last equation is due to Step 2. □

Proof of Corollary 3.3. Proof of (3.12). We observe

$$\limsup_{n\to\infty}\sup_{h\in\mathcal{H}_0} \mathrm{P}_h\left(h \notin \mathcal{C}_n(\alpha)\right) = \limsup_{n\to\infty}\sup_{h\in\mathcal{H}_0} \mathrm{P}_h\left(\max_{J\in\mathcal{I}_n}\frac{n\widehat{D}_J(h)}{\widehat{\eta}_J\,\widehat{\mathrm{v}}_J(h)} > \sqrt{2}\right) \leq \alpha,$$

where the last inequality is due to step 1 of the proof of Theorem 3.3 and step 3 of the proof of Theorem 3.4.

Proof of (3.13). Let J^* be as be as in step 2 of the proof of Theorem 3.3. We observe for all $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ that

$$\begin{aligned}
\mathrm{P}_h\left(h \notin \mathcal{C}_n(\alpha)\right) &= \mathrm{P}_h\left(\max_{J\in\mathcal{I}_n} \frac{n\widehat{D}_J(h)}{\widehat{\eta}_J \widehat{\mathrm{v}}_J(h)} > \sqrt{2}\right)\\
&= 1 - \mathrm{P}_h\left(\max_{J\in\mathcal{I}_n} \frac{n\widehat{D}_J(h)}{\widehat{\eta}_J \,\widehat{\mathrm{v}}_J(h)} \leq \sqrt{2}\right)\\
&\geq 1 - \mathrm{P}_h\left(\frac{n\widehat{D}_{J^*}(h)}{\widehat{\eta}_{J^*} \,\widehat{\mathrm{v}}_{J^*}(h)} \leq \sqrt{2}\right)\\
&\geq 1-\alpha,
\end{aligned}$$

for n sufficiently large, where the last inequality is due to step 2 of the proof of Theorem 3.3 and step 3 of the proof of Theorem 3.4. □

Proof of Corollary 3.4. For any $h \in \mathcal{H}$, we analyze the diameter of the confidence set $\mathcal{C}_n(\alpha)$ under P_h for some For all $h_1 \in \mathcal{C}_n(\alpha) \subset \mathcal{H}$ it holds for all $J \in \mathcal{I}_n$ by using the definition of the projection Q_J:

$$\begin{aligned}
\|h-h_1\|_\mu &\leq \|Q_J\Pi_J(h-h_1)\|_\mu + \|\Pi_J h - h\|_\mu + \|\Pi_J h_1 - h_1\|_\mu\\
&\leq \|Q_J(h-h_1)\|_\mu + O(J^{-p/d_x}),
\end{aligned}$$

where the second inequality due to the triangular inequality and Assumptions 2 (ii) and 3 (ii). The upper bound established in (F.4) yields:

$$\left|\|Q_J(h-h_1)\|_\mu^2 - \widehat{D}_J(h_1)\right| \leq n^{-1/2}\left\|\langle Q_J(h-h_1), \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\right\| + n^{-1}\mathrm{v}_J$$

with probability approaching one. Consequently, the definition of the confidence set $\mathcal{C}_n(\alpha)$ with $h_1 \in \mathcal{C}_n(\alpha)$ gives

$$\begin{aligned}
\|Q_J(h-h_1)\|_\mu^2 &\leq \widehat{D}_J(h_1) + n^{-1/2}\left\|\langle Q_J(h-h_1), \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\right\| + n^{-1}\mathrm{v}_J\\
&\leq \sqrt{2}n^{-1}\widehat{\eta}_J\,\widehat{\mathrm{v}}_J + n^{-1/2}\left\|\langle Q_J(h-h_1), \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\right\| + n^{-1}\mathrm{v}_J\\
&\leq C\Big(n^{-1}\eta_J\,\mathrm{v}_J + n^{-1/2}\left\|\langle Q_J(h-h_1), \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\right\|\Big)
\end{aligned}$$

with probability approaching one. We may choose $J = J_0 \in \mathcal{I}_n$ for n sufficiently large and hence, the result follows by applying Lemma F.1 and Assumption 5 (ii), i.e., $\eta_J = O(\sqrt{\log\log n})$. □

E. Proofs of Adaptive IT Results in Subsection 4.2

For the adaptive IT results, we may consider restriction to the index set $\mathcal{I}_n$ given in (B.1), but where the uppder bound $\overline{J}$ is replaced by $\overline{K}$. As in Appendix B this implies that $\widehat{\eta}_K(\alpha)$ is deterministic and is denoted by η_K in the following.

Proof of Theorem 4.1. The proof is based on two steps, where we bound the first and second type error of $\widehat{\mathtt{IT}}_n$ separately.

Step 1: The definition of the test statistic $\widehat{\mathtt{IT}}_n$ implies for all $h \in \mathcal{H}_0$:

$$\begin{aligned}
\mathrm{P}_h\Big(\widehat{\mathtt{IT}}_n = 1\Big) \leq & \underbrace{\mathrm{P}_h\Big(\max_{K\in\mathcal{I}_n} \frac{n\,\widehat{D}_K(h^{\mathrm{R}})}{\eta_K\,\widehat{\mathrm{v}}_K(h^{\mathrm{R}})} > \sqrt{1/2}\Big)}_{I} \\
& + \underbrace{\mathrm{P}_h\Big(n \max_{K\in\mathcal{I}_n}\sup_{h\in\mathcal{H}^{\mathrm{R}}} \big|\widehat{D}_K(h) - \widehat{D}_K(h^{\mathrm{R}})\big|/(\eta_K\,\widehat{\mathrm{v}}_K(h^{\mathrm{R}})) > \sqrt{1/8}\Big)}_{II} \\
& + \underbrace{\mathrm{P}_h\Big(\max_{K\in\mathcal{I}_n}\Big|\frac{n\,\widehat{D}_K(\widehat{h}^{\mathrm{R}})}{\eta_K\,\widehat{\mathrm{v}}_K(h^{\mathrm{R}})}\Big| \max_{K\in\mathcal{I}_n}\Big|1 - \frac{\widehat{\mathrm{v}}_K(h^{\mathrm{R}})}{\widehat{\mathrm{v}}_K(\widehat{h}^{\mathrm{R}})}\Big| > \sqrt{1/8}\Big)}_{III}.
\end{aligned}$$

We obtain $I = o(1)$ following the proof of Theorem 3.1 (with U_i replaced by $\rho(Y_i, h^{\mathrm{R}}(X_i))$ and $\widehat{A}'G\widehat{A}$ replaced by $(B'B/n)^-$). Consider II. The definition of the estimator $\widehat{D}_K$ implies for all $K \in \mathcal{I}_n$ and $h \in \mathcal{H}^{\mathrm{R}}$ that

$$\begin{aligned}
& \frac{1}{\eta_K\,\mathrm{v}_K(n-1)} \sum_{i\neq i'} \big(\rho(Y_i, h(X_i)) - \rho(Y_i, h^{\mathrm{R}}(X_i))\big)\big(\rho(Y_{i'}, h(X_{i'})) - \rho(Y_{i'}, h^{\mathrm{R}}(X_{i'}))\big)\widetilde{b}^K(W_i)'\widetilde{b}^K(W_{i'}) \\
& + \frac{1}{\eta_K\,\mathrm{v}_K(n-1)} \sum_{i\neq i'} \rho(Y_i, h^{\mathrm{R}}(X_i))\big(\rho(Y_{i'}, h(X_{i'})) - \rho(Y_{i'}, h^{\mathrm{R}}(X_{i'}))\big)\widetilde{b}^K(W_i)'\widetilde{b}^K(W_{i'}) \\
& = T_{1,K}(h) + T_{2,K}(h).
\end{aligned}$$

Consider $T_{1,K}(h)$. We obtain

$$\begin{aligned}
|T_{1,K}(h)| \leq & \frac{1}{\eta_K\,\mathrm{v}_K}\Big\|n^{-1/2}\sum_{i=1}^n \big(\rho(Y_i, h(X_i)) - \rho(Y_i, h^{\mathrm{R}}(X_i))\big)\widetilde{b}^K(W_i)\Big\|^2 \\
& + \frac{1}{n\eta_K\,\mathrm{v}_K}\sum_{i=1}^n \Big\|\big(\rho(Y_i, h(X_i)) - \rho(Y_i, h^{\mathrm{R}}(X_i))\big)\widetilde{b}^K(W_i)\Big\|^2 \\
= & \, T_{11,K}(h) + T_{12,K}(h).
\end{aligned}$$

Consider $T_{11,K}(h)$. We obtain

$$\begin{aligned}
&\mathrm{P}_h\Big(\max_{K\in\mathcal{I}_n}\sup_{h\in\mathcal{H}^{\mathrm{R}}} T_{11,K}(h) > 1\Big)\\
&\leq \mathrm{P}_h\Big(\exists K\in\mathcal{I}_n : \sup_{h\in\mathcal{H}^{\mathrm{R}}}\Big\|\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big(\big(\rho(Y_i,h(X_i))-\rho(Y_i,h^{\mathrm{R}}(X_i))\big)\widetilde{b}^K(W_i)-\mathrm{E}\left[\big(\rho(Y,h(X))-\rho(Y,h^{\mathrm{R}}(X))\big)\widetilde{b}^K(W)\right]\Big)\Big\|^2\\
&\qquad\qquad > \frac{\eta_K \mathrm{v}_K}{2}-\sup_{h\in\mathcal{H}^{\mathrm{R}}}\big\|\mathrm{E}\left[\big(\rho(Y,h(X))-\rho(Y,h^{\mathrm{R}}(X))\big)\widetilde{b}^K(W)\right]\big\|^2\Big).
\end{aligned}$$

We evaluate by using Assumption 3 (ii) that uniformly in $h\in\mathcal{H}^{\mathrm{R}}$:

$$\big\|\mathrm{E}\left[\big(\rho(Y,h(X))-\rho(Y,h^{\mathrm{R}}(X))\big)\widetilde{b}^K(W)\right]\big\| \leq \|h-h^{\mathrm{R}}\|_\mu \leq \Delta_n.$$

Now using that for each $K\in\mathcal{I}_n$ and $1\leq k\leq K$ we have

$$\mathrm{E}\sup_{h\in\mathcal{H}^{\mathrm{R}}}\big|\big(\rho(Y,h(X))-\rho(Y,h^{\mathrm{R}}(X))\big)\widetilde{b}_k(W)|^2 \leq C\Delta_n^{2\kappa}$$

by Assumption 6 (i). Consequently,following the derivation of the upper bound (D.1) we obtain

$$\begin{aligned}
\mathrm{P}_h\Big(\max_{K\in\mathcal{I}_n(\Delta_n)}\sup_{h\in\mathcal{H}^{\mathrm{R}}} T_{11,K}(h) > 1\Big) &\leq \max_{K\in\mathcal{I}_n}\frac{K\Delta_n^{2\kappa}}{\eta_K\mathrm{v}_K-\Delta_n^2}\left(\int_0^1\sqrt{1+\log N_{[]}(wC,\mathcal{H}^{\mathrm{R}},\|\cdot\|_2)}dw\right)^2\\
&\leq (1-c_0)^{-1}\max_{K\in\mathcal{I}_n}\frac{K\Delta_n^{2\kappa}}{\eta_K\mathrm{v}_K} = o(1)
\end{aligned}$$

where the second inequality is based on the inequality $\Delta_n^{2\kappa}\leq c_0\eta_K\mathrm{v}_K$ for some $0<c_0<1$ and all $K\in\mathcal{I}_n$ as n becomes sufficiently large. Consider $T_{12,K}(h)$. We obtain

$$\max_{K\in\mathcal{I}_n}\sup_{h\in\mathcal{H}^{\mathrm{R}}} T_{12,K}(h) \leq \max_{J\in\mathcal{I}_n}\frac{\big(\Delta_n\zeta_K\big)^2}{\eta_K\mathrm{v}_K} = o(1).$$

The bound for $T_{2,K}(h)$ follows analogously.

Step 2: It is sufficient to control the second type error of the statistic

$$\widetilde{\mathtt{IT}}_n = \mathbb{1}\left\{\max_{K\in\mathcal{I}_n}|n\widehat{D}_K(\widehat{h}^{\mathrm{R}})/(\eta''_K\mathrm{v}_K)|>1\right\}$$

for some $\eta''_K>0$. Let K^* denote the largest integer such that $n^{-1}\sqrt{K^*}\leq\mathsf{r}_n^2$. We evaluate for all $h\in\mathcal{M}_1(\delta^\circ,\mathsf{r}_n)$ that

$$\begin{aligned}
\mathrm{P}_h\left(\widetilde{\mathtt{IT}}_n=0\right) &= \mathrm{P}_h\left(n\,\widehat{D}_K(\widehat{h}^{\mathrm{R}})\leq\eta''_K\,\mathrm{v}_K \text{ for all } K\in\mathcal{I}_n\right)\\
&\leq \mathrm{P}_h\left(n\,\widehat{D}_{K^*}(\widehat{h}^{\mathrm{R}})\leq\eta''_{K^*}\,\mathrm{v}_{K^*}\right).
\end{aligned}$$

We make use of the notation $B_K=\|\mathrm{E}_h[V^K]\|_2^2-\|m(\cdot,h^{\mathrm{R}})\|^2_{L^2(W)}$ where we write $V^K=$

$\rho(Y, h^{\mathrm{R}}(X))G_b^{-1/2}b^{K^*}(W)$. We obtain uniformly over $h \in \mathcal{M}_1(\delta^\circ, \mathsf{r}_n)$ that

$$\begin{aligned}
&\mathrm{P}_h\left(n\,\widehat{D}_{K^*}(\widehat{h}^{\mathrm{R}}) \leq \eta''_{K^*}\,\mathrm{v}_{K^*}\right) = \mathrm{P}_h\left(\|\,\mathrm{E}_h[V^{K^*}]\|_2^2 - \widehat{D}_{K^*}(\widehat{h}^{\mathrm{R}}) > \|\,\mathrm{E}_h[V^{K^*}]\|_2^2 - \frac{\eta''_{K^*}\,\mathrm{v}_{K^*}}{n}\right)\\
&\leq \underbrace{\mathrm{P}_h\left(\Big|\frac{8}{n^2}\sum_{k=1}^{K^*}\sum_{i<i'}\big(V_{ik}V_{i'k} - \mathrm{E}_h[V_{1k}]^2\big)\Big| > \mathrm{E}[m^2(W, h^{\mathrm{R}})] - \frac{\eta''_{K^*}\,\mathrm{v}_{K^*}}{n} + B_{K^*}\right)}_{IV}\\
&+ \underbrace{\mathrm{P}_h\left(\Big|\frac{8}{n^2}\sum_{i<i'}\rho(Y_i, h^{\mathrm{R}}(X_i))\rho(Y_{i'}, h^{\mathrm{R}}(X_{i'}))b^{K^*}(W_i)'\big(\widehat{G}_b^{-1} - G_b^{-1}\big)b^{K^*}(W_{i'})\Big| > \mathrm{E}[m^2(W, h^{\mathrm{R}})] - \frac{\eta''_{K^*}\,\mathrm{v}_{K^*}}{n} + B_{K^*}\right)}_{V}\\
&+ \underbrace{\mathrm{P}_h\left(\big|\widehat{D}_{K^*}(\widehat{h}^{\mathrm{R}}) - \widehat{D}_{K^*}(h^{\mathrm{R}})\big| > \mathrm{E}[m^2(W, h^{\mathrm{R}})] - \frac{\eta''_{K^*}\,\mathrm{v}_{K^*}}{n} + B_{K^*}\right)}_{VI},
\end{aligned}$$

using $n/(n-1) \leq 2$ for $n \geq 2$. Consider IV. We observe $|B_{K^*}| \leq C_B\,\mathsf{r}_n^2$ which is due to the upper bound

$$\|\,\mathrm{E}_h[V^{K^*}]\| = \big(\mathrm{E}_h[\rho(Y, h^{\mathrm{R}}(X))b^{K^*}(W)']G_b^{-1}\,\mathrm{E}_h[\rho(Y, h^{\mathrm{R}}(X))b^{K^*}(W)]\big)^{1/2} = \|\Pi_{K^*}m(\cdot, h^{\mathrm{R}})\|_{L^2(W)}.$$

Further, from the proof of Lemma F.9 we deduce

$$\mathrm{E}_h\left|\frac{1}{n(n-1)}\sum_{k=1}^{K^*}\sum_{i\neq i'}\big(V_{ik}V_{i'k} - \mathrm{E}_h[V_{1k}]^2\big)\right|^2 \leq \frac{C}{n}\left(\|\Pi_{K^*}m(\cdot, h^{\mathrm{R}})\|_{L^2(W)}^2 + \frac{\mathrm{v}_{K^*}^2}{n}\right).$$

Consequently, Markov's inequality yields

$$IV = O\left(\frac{n^{-1}\|\Pi_{K^*}m(\cdot, h^{\mathrm{R}})\|_{L^2(W)}^2 + n^{-2}\mathrm{v}_{K^*}^2}{\big(\|m(\cdot, h^{\mathrm{R}})\|_{L^2(W)}^2 - \eta''_{K^*}\,n^{-1}\mathrm{v}_{K^*} + B_{K^*}\big)^2}\right) \tag{E.1}$$

In the following, we distinguish between two cases. First, consider the case where $n^{-2}\mathrm{v}_{K^*}^2$ dominates the summand in the numerator. For any $h \in \mathcal{M}_1(\delta^\circ, \mathsf{r}_n)$ we have $\|m(\cdot, h^{\mathrm{R}})\|_{L^2(W)}^2 \geq \delta^\circ\,\mathsf{r}_n^2$ and hence, using that $\mathrm{v}_{K^*} \geq c_0\sqrt{K^*}$ for some constant $0 < c_0 < 1$ we obtain the lower bound

$$\|m(\cdot, h^{\mathrm{R}})\|_{L^2(W)}^2 - \eta''_{K^*}\,n^{-1}\mathrm{v}_{K^*} + B_{K^*} \geq (\delta^\circ - c_0 - C_B)\,\mathsf{r}_n^2 \geq C_0\mathsf{r}_n^2$$

for some constant $C_0 := \delta^\circ - c_0 - C_B$ which is positive for any $\delta^\circ > c_0 + C_B$. From inequality (E.1) we infer $IV = O(\mathsf{r}_n^{-4}n^{-2}\mathrm{v}_{K^*}^2) = o(1)$. Second, consider the case where $n^{-1}\|\Pi_{K^*}m(\cdot, h^{\mathrm{R}})\|_{L^2(W)}$ dominates. For any $h \in \mathcal{M}_1(\delta^\circ, \mathsf{r}_n)$ we have $\|m(\cdot, h^{\mathrm{R}})\|_{L^2(W)}^2 \geq$

$\delta^\circ \mathsf{r}_n^2 > \delta^\circ n^{-1}\sqrt{K^*}$ and hence, obtain the lower bound

$$\begin{aligned}\|m(\cdot,h^{\mathrm{R}})\|_{L^2(W)}^2 - \eta''_{K^*}\, n^{-1}\mathrm{v}_{K^*} + B_{K^*} &\geq \left(1-\frac{1}{\delta^\circ}-\frac{C_B}{\delta^\circ}\right)\|m(\cdot,h^{\mathrm{R}})\|_{L^2(W)}^2\\ &\geq c_1\,\|m(\cdot,h^{\mathrm{R}})\|_{L^2(W)}^2\end{aligned}$$

for some constant $c_1 := 1-(1+C_B)/\delta^\circ$ which is positive for any $\delta^\circ > 1+C_B$. Hence, inequality (E.1) yields for all $h \in \mathcal{M}_1(\delta^\circ, \mathsf{r}_n)$ that

$$IV = O\Big(n^{-1}\frac{1}{\|m(\cdot,h^{\mathrm{R}})\|_{L^2(W)}^2}\Big) = O\left(n^{-1}\mathsf{r}_n^{-2}\right) = o(1).$$

Finally, $V = o(1)$ by making use of Lemma F.4 (with U_i replaced by $\rho(Y_i, h^{\mathrm{R}}(X_i))$ and $\widehat{A}'G\widehat{A}$ replaced by $(B'B/n)^-$) and $VI = o(1)$ by following step 1. □

F. Technical Results

Theorem F.1. *Let Assumptions 1–3 be satisfied. Then, it holds*

$$\big|\widehat{D}_J - \|h-h_0\|_\mu^2\big| = O_p\left(n^{-1}s_J^{-2}\sqrt{J} + n^{-1/2}\big\|\langle Q_J(h-h_0),\psi^J\rangle_\mu'(G^{-1/2}S)_l^-\big\| + J^{-2p/d_x}\right).$$

Proof. We make use of the decomposition

$$\widehat{D}_J - \|h-h_0\|_\mu^2 = \widehat{D}_J - \|Q_J(h-h_0)\|_\mu^2 + \|Q_J(h-h_0)\|_\mu^2 - \|h-h_0\|_\mu^2.$$

Note that

$$\begin{aligned}\|Q_J(h-h_0)\|_\mu^2 &= \int \left(\psi^J(x)'A\,\mathrm{E}_h[(Y-h_0(X))b^K(W)]\right)^2\mu(x)dx\\ &= \mathrm{E}_h[(Y-h_0(X))b^K(W)]'A'\underbrace{\int \psi^J(x)\psi^J(x)'\mu(x)dx}_{=G}\,A\,\mathrm{E}_h[(Y-h_0(X))b^K(W)]\\ &= \big\|G^{1/2}A\,\mathrm{E}_h[(Y-h_0(X))b^K(W)]\big\|^2\\ &= \|\,\mathrm{E}_h[V^J]\|^2\end{aligned}$$

using the notation $V_i^J = (Y_i - h_0(X_i))G^{1/2}Ab^K(W_i)$. Thus, the definition of the estimator

$\widehat{D}_J$ implies

$$
\begin{aligned}
&\widehat{D}_J - \|Q_J(h-h_0)\|_\mu^2 \\
&= \frac{1}{n(n-1)} \sum_{j=1}^{J} \sum_{i \neq i'} \big(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\big) && \text{(F.1)} \\
&\quad + \frac{1}{n(n-1)} \sum_{i\neq i'} Y_i Y_{i'} b^K(W_i)' \Big(A'GA - \widehat{A}'G\widehat{A}\Big) b^K(W_{i'}), && \text{(F.2)}
\end{aligned}
$$

where we bound both summands on the right hand side separately in the following. Consider the summand in (F.1), we observe

$$
\Big|\sum_{j=1}^{J}\sum_{i\neq i'}\big(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\big)\Big|^2 = \sum_{j,j'=1}^{J}\sum_{i\neq i'}\sum_{i''\neq i'''}\big(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\big)\big(V_{i''j'}V_{i'''j'} - \mathrm{E}_h[V_{1j'}]^2\big)
$$

We distinguish three different cases. First: i, i', i'', i''' are all different, second: either $i = i''$ or $i' = i'''$, or third: $i = i'$ and $i' = i'''$. We thus calculate for each $j, j' \geq 1$ that

$$
\begin{aligned}
&\sum_{i\neq i'}\sum_{i''\neq i'''}\big(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\big)\big(V_{i''j'}V_{i'''j'} - \mathrm{E}_h[V_{1j'}]^2\big) \\
&\qquad = \sum_{i,i',i'',i''' \text{all different}}\big(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\big)\big(V_{i''j'}V_{i'''j'} - \mathrm{E}_h[V_{1j'}]^2\big) \\
&\qquad\quad + 2\sum_{i\neq i'\neq i''}\big(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\big)\big(V_{i''j'}V_{i'j'} - \mathrm{E}_h[V_{1j'}]^2\big) \\
&\qquad\quad + \sum_{i\neq i'}\big(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\big)\big(V_{ij'}V_{i'j'} - \mathrm{E}_h[V_{1j'}]^2\big).
\end{aligned}
$$

Due to independent observations we have

$$
\sum_{i,i',i'',i''' \text{all different}} \mathrm{E}_h\Big[\big(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\big)\big(V_{i''j'}V_{i'''j'} - \mathrm{E}_h[V_{1j'}]^2\big)\Big] = 0
$$

Consequently, we calculate

$$
\begin{aligned}
&\mathrm{E}_h \Big| \sum_{j=1}^{J} \sum_{i \neq i'} \big(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\big) \Big|^2 \\
&= 2n(n-1)(n-2) \underbrace{\sum_{j,j'=1}^{J} \mathrm{E}_h\Big[\big(V_{1j}V_{2j} - \mathrm{E}_h[V_{1j}]^2\big)\big(V_{3j'}V_{2j'} - \mathrm{E}_h[V_{1j'}]^2\big)\Big]}_{I} \\
&+ n(n-1) \underbrace{\sum_{j,j'=1}^{J} \mathrm{E}_h\Big[\big(V_{1j}V_{2j} - \mathrm{E}_h[V_{1j}]^2\big)\big(V_{1j'}V_{2j'} - \mathrm{E}_h[V_{1j'}]^2\big)\Big]}_{II}.
\end{aligned}
$$

To bound the summand I we observe that

$$
\begin{aligned}
I &= \sum_{j,j'=1}^{J} \mathrm{E}_h[V_{1j}]\, \mathrm{E}_h[V_{1j'}]\, \mathbb{C}ov_h(V_{1j}, V_{1j'}) \\
&= \mathrm{E}_h[V_1^J]'\, \mathbb{C}ov_h(V_1^J, V_1^J)\, \mathrm{E}_h[V_1^J] \\
&\leq \lambda_{\max}\big(\mathbb{V}ar_h((Y - h_0(X))G_b^{-1/2}b^K(W))\big) \big\| G_b^{1/2} A' G^{1/2}\, \mathrm{E}_h[V_1^J] \big\|^2 \\
&\leq \overline{\sigma}^2 \big\| \mathrm{E}_h[(Y - h_0(X))b^K(W)]' A' G A G_b^{1/2} \big\|^2 \\
&= \overline{\sigma}^2 \Big\| \int Q_J(h - h_0)(x) \psi^J(x)' \mu(dx) (G_b^{-1/2} S)_l^- \Big\|^2
\end{aligned}
$$

by using the notation $V_i^J = (Y_i - h_0(X_i))G^{1/2} A b^K(W_i)$, $AG_b^{1/2} = (G_b^{-1/2}S)_l^-$, and Lemma F.8, i.e., $\lambda_{\max}\big(\mathbb{V}ar_h((Y - h_0(X))\widetilde{b}^K(W))\big) \leq \overline{\sigma}^2$. Consider II. We observe

$$
\begin{aligned}
II &= n(n-1) \sum_{j,j'=1}^{J} \mathrm{E}_h[V_{1j}V_{1j'}]^2 - n(n-1)\Big(\sum_{j=1}^{J} \mathrm{E}_h[V_{1j}]^2\Big)^2 \\
&\leq n(n-1) \sum_{j,j'=1}^{J} \mathrm{E}_h[V_{1j}V_{1j'}]^2 \\
&= n(n-1) \mathrm{v}_J^2.
\end{aligned}
$$

The upper bounds derived for the terms I and II imply for all $n \geq 2$:

$$
\begin{aligned}
&\mathrm{E}_h \Big| \frac{1}{n(n-1)} \sum_{j=1}^{J} \sum_{i \neq i'} \big(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\big) \Big|^2 \\
&\qquad \leq 2\overline{\sigma}^2 \left(\frac{1}{n} \big\| \langle Q_J(h - h_0), \psi^J \rangle_\mu' (G_b^{-1/2} S)_l^- \big\|^2 + \frac{\mathrm{v}_J^2}{n^2} \right). \qquad \text{(F.3)}
\end{aligned}
$$

Consequently, equality (F.2) together with Lemma F.4 yields

$$\widehat{D}_J - \|Q_J(h-h_0)\|_\mu^2 = O_p\left(n^{-1/2}\big\|\langle Q_J(h-h_0), \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\big\| + n^{-1}v_J^1\right), \tag{F.4}$$

which implies the variance part by employing Lemma F.1. Finally, Lemma F.3 implies for the bias term

$$\|Q_J(h-h_0)\|_\mu^2 - \|h-h_0\|_\mu^2 = O(J^{-2p/d})$$

which completes the proof. □

Lemma F.1. *Let Assumption 1 (ii) be satisfied. Then, it holds*

$$v_J \le \overline{\sigma}^2 s_J^{-2}\sqrt{J}.$$

Proof. In the following, let e_j be the unit vector with 1 at the j–th position. We obtain

$$\begin{aligned}
v_J^2 &= \sum_{j,j'=1}^J \mathrm{E}_h[V_{1j}V_{1j'}]^2 \\
&= \sum_{j=1}^J \Big\| \mathrm{E}[\underbrace{\mathrm{E}_h[(Y-h_0(X))^2|W]e_j'G^{1/2}Ab^K(W)}_{=:\chi_j(W)}G^{1/2}Ab^K(W)]\Big\|^2 \\
&= \sum_{j=1}^J \Big\|G^{1/2}AG_b^{1/2}\,\mathrm{E}[\chi_j(W)\widetilde{b}^K(W)]\Big\|^2 \\
&\le s_J^{-2}\sum_{j=1}^J \big\|\,\mathrm{E}[\chi_j(W)\widetilde{b}^K(W)]\big\|^2,
\end{aligned}$$

using the relationship $\|G^{1/2}AG_b^{1/2}\| = s_J^{-1}$. For all $j \ge 1$ we have the upper bound

$$\big\|\,\mathrm{E}[\chi_j(W)\widetilde{b}^K(W)]\big\| \le \|\chi_j\|_{L^2(W)}.$$

Now using that $\sup_{w\in\mathcal{W}}\sup_{h\in\mathcal{H}}\mathrm{E}_h[(Y-h_0(X))^2|W=w] \le \overline{\sigma}^2$ due to Assumption 1 (ii), we get

$$\begin{aligned}
\sum_{j=1}^J \|\chi_j\|_{L^2(W)}^2 &\le \overline{\sigma}^4 \sum_{j=1}^J \mathrm{E}\,|e_j'G^{1/2}Ab^K(W)|^2 \\
&= \overline{\sigma}^4 \sum_{j=1}^J e_j'G^{1/2}AG_bA'G^{1/2}e_j \\
&\le \overline{\sigma}^4 s_J^{-2} J,
\end{aligned}$$

which implies the assertion. ∎

Lemma F.2. *Let Assumption 1 (iii) be satisfied. Then, it holds*

$$\sqrt{\sum_{j=1}^{J} s_j^{-4}} \leq \underline{\sigma}^{-2} \mathrm{v}_J,$$

where s_j^{-1}, $1 \leq j \leq J$, are the nondecreasing singular values of $G^{1/2}AG_b^{1/2}$.

Proof. In the following, let e_j be the unit vector with 1 at the j–th position. Introduce a unitary matrix Q such that by Schur decomposition

$$Q'G^{1/2}AG_bA'G^{1/2}Q = \operatorname{diag}(s_1^{-2}, \ldots, s_J^{-2}).$$

We make use of the notation $\widetilde{V}_i^J = (Y_i - h_0(X_i))Q'G^{1/2}Ab^K(W_i)$. Now since the Frobenius norm is invariant under unitary matrix multiplication we have

$$\begin{aligned}
\mathrm{v}_J^2 &= \sum_{j,j'=1}^{J} \mathrm{E}_h[\widetilde{V}_{1j}\widetilde{V}_{1j'}]^2 \\
&\geq \sum_{j=1}^{J} \mathrm{E}_h[\widetilde{V}_{1j}^2]^2 \\
&= \sum_{j=1}^{J} \big(\mathrm{E}\,|(Y - h_0(X))e_j'Q'G^{1/2}Ab^K(W)|^2\big)^2 \\
&\geq \underline{\sigma}^4 \sum_{j=1}^{J} \big(\mathrm{E}[e_j'Q'G^{1/2}Ab^K(W)b^K(W)'A'G^{1/2}Qe_j]\big)^2 \\
&= \underline{\sigma}^4 \sum_{j=1}^{J} \big(e_j'Q'G^{1/2}AG_bA'G^{1/2}Qe_j\big)^2 \\
&= \underline{\sigma}^4 \sum_{j=1}^{J} \big(e_j' \operatorname{diag}(s_1^{-2}, \ldots, s_J^{-2})e_j\big)^2 \\
&\geq \underline{\sigma}^4 \sum_{j=1}^{J} s_j^{-4},
\end{aligned}$$

using $\inf_{w\in\mathcal{W}} \inf_{h\in\mathcal{H}} \mathrm{E}_h[(Y - h_0(X))^2|W = w] \geq \underline{\sigma}^2$ by Assumption 1 (iii). ∎

Lemma F.3. *Let Assumptions 2 and 3 be satisfied. Then, for all $h \in \mathcal{H}$ we have*

$$\|Q_J(h - h_0)\|_\mu = \|h - h_0\|_\mu + O\big(J^{-p/d}\big).$$

Proof. Using the notation $\widetilde{b}^K(\cdot) := G_b^{-1/2} b^K(\cdot)$, we observe for all $h \in \mathcal{H}$ that

$$
\begin{aligned}
\|Q_J(h-h_0)\|_\mu &= \big\|(G_b^{-1/2} S G^{-1/2})_l^- \operatorname{E}[\widetilde{b}^K(W)(h-h_0)(X)]\big\| \\
&\le \big\|(G_b^{-1/2} S G^{-1/2})_l^- \operatorname{E}[\widetilde{b}^K(W)(\Pi_J h - \Pi_J h_0)(X)]\big\| \\
&\quad + \big\|(G_b^{-1/2} S G^{-1/2})_l^- \operatorname{E}[\widetilde{b}^K(W)((h-h_0)(X) - (\Pi_J h - \Pi_J h_0)(X))]\big\| \\
&\le \|\Pi_J h - \Pi_J h_0\|_\mu + s_J^{-1} \|\Pi_K T((h-h_0) - (\Pi_J h - \Pi_J h_0))\|_{L^2(W)} \\
&\le \|\Pi_J h - \Pi_J h_0\|_\mu + O\big(J^{-p/d}\big)
\end{aligned}
$$

by making use of Assumption 3 (ii). □

Lemma F.4. *Let Assumptions 1–3 be satisfied. Then, uniformly in $h \in \mathcal{H}$ it holds*

$$
\begin{aligned}
&\frac{1}{n(n-1)} \sum_{i \neq i'} \big(Y_i - h_0(X_i)\big)\big(Y_{i'} - h_0(X_{i'})\big) b^K(W_i)' \Big(A'GA - \widehat{A}'G\widehat{A}\Big) b^K(W_{i'}) \\
&\qquad = O_p\Big(n^{-1} s_J^{-2} \sqrt{J} + n^{-1/2} \big\|\langle Q_J(h-h_0), \psi^J\rangle_\mu' (G_b^{-1/2} S)_l^- \big\|\Big).
\end{aligned}
$$

Proof. In the proof, we establish an upper bound of

$$
\begin{aligned}
&\frac{1}{n^2} \sum_{i,i'} \big(Y_i - h_0(X_i)\big)\big(Y_{i'} - h_0(X_{i'})\big) b^K(W_i)' \Big(A'GA - \widehat{A}'G\widehat{A}\Big) b^K(W_{i'}) \\
&= \operatorname{E}[(h-h_0)(X) b^K(W)]' \Big(A'GA - \widehat{A}'G\widehat{A}\Big) \operatorname{E}[(h-h_0)(X) b^K(W)] \\
&+ 2\Big(\frac{1}{n}\sum_i \big(Y_i - h_0(X_i)\big) b^K(W_i)' - \operatorname{E}[(h-h_0)(X) b^K(W)]'\Big) \Big(A'GA - \widehat{A}'G\widehat{A}\Big) \\
&\qquad \times \operatorname{E}[(h-h_0)(X) b^K(W)] \\
&+ \Big(\frac{1}{n}\sum_i \big(Y_i - h_0(X_i)\big) b^K(W_i)' - \operatorname{E}[(h-h_0)(X) b^K(W)]'\Big) \Big(A'GA - \widehat{A}'G\widehat{A}\Big) \\
&\qquad \times \Big(\frac{1}{n}\sum_i \big(Y_i - h_0(X_i)\big) b^K(W_i)' - \operatorname{E}[(h-h_0)(X) b^K(W)]'\Big)
\end{aligned}
$$

uniformly in $h \in \mathcal{H}$. It is sufficient to bound the first summand on the right hand side. We make use of the decomposition

$$
\begin{aligned}
&\operatorname{E}[(h-h_0)(X) b^K(W)]' \Big(A'GA - \widehat{A}'G\widehat{A}\Big) \operatorname{E}[(h-h_0)(X) b^K(W)] \\
&\qquad = 2 \operatorname{E}[(h-h_0)(X) b^K(W)]' A'G(A - \widehat{A}) \operatorname{E}[(h-h_0)(X) b^K(W)] \\
&\qquad - \operatorname{E}[(h-h_0)(X) b^K(W)]' (A - \widehat{A})' G (A - \widehat{A}) \operatorname{E}[(h-h_0)(X) b^K(W)] \\
&\qquad = 2T_1 - T_2,
\end{aligned}
$$

where we bound each summand separately in what follows. Consider T_1. Below, we show the result

$$T_1 = O_p\Big(n^{-1/2}\big\|\langle Q_J(h-h_0), \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\big\|\Big). \tag{F.5}$$

To do so, we make use of the decomposition

$$\begin{aligned} T_1 &= \mathrm{E}[(h-h_0)(X)b^K(W)]'A'G(\widehat{A}-A)\,\mathrm{E}[\Pi_J(h-h_0)(X)b^K(W)] \\ &\quad + \mathrm{E}[(h-h_0)(X)b^K(W)]'A'G(\widehat{A}-A)\,\mathrm{E}[(h-h_0-\Pi_J(h-h_0))(X)b^K(W)]. \end{aligned} \tag{F.6}$$

Consider the first summand on the right hand side of the equation. Using the definition of the left pseudo inverse we can write $\widehat{A} = (\widehat{G}_b^{-1/2}\widehat{S})_l^-\widehat{G}_b^{-1/2}$ where $\widehat{S} = n^{-1}\sum_i b^K(W_i)\psi^J(X_i)'$. Making use of the relation $Q_J\Pi_J h = \Pi_J h$ and $\widehat{S}G^{-1}\langle h, \psi^J\rangle_\mu = n^{-1}\sum_i \Pi_J h(X_i)b^K(W_i)$ yields

$$\begin{aligned} &\mathrm{E}[(h-h_0)(X)b^K(W)]'A'G(A-\widehat{A})\,\mathrm{E}[\Pi_J(h-h_0)(X)b^K(W)] \\ &= \int Q_J(h-h_0)(x)\Big(\Pi_J(h-h_0)(x) - \psi^J(x)'(\widehat{G}_b^{-1/2}\widehat{S})_l^-\widehat{G}_b^{-1/2}\,\mathrm{E}[(h-h_0)(X)b^K(W)]\Big)\mu(x)dx \\ &= \langle Q_J(h-h_0), \psi^J\rangle_\mu'(\widehat{G}_b^{-1/2}\widehat{S})_l^-\widehat{G}_b^{-1/2}\Big(\frac{1}{n}\sum_i \Pi_J(h-h_0)(X_i)b^K(W_i) - \mathrm{E}[(h-h_0)(X)b^K(W)]\Big) \\ &= \langle Q_J(h-h_0), \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\Big(\frac{1}{n}\sum_i \Pi_J(h-h_0)(X_i)\widetilde{b}^K(W_i) - \mathrm{E}[\Pi_J(h-h_0)(X)\widetilde{b}^K(W)]\Big) \\ &\quad + \langle Q_J(h-h_0), \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-G_b^{-1/2}S'\Big((\widehat{G}_b^{-1/2}\widehat{S})_l^-\widehat{G}_b^{-1/2}G_b^{1/2} - (G_b^{-1/2}S)_l^-\Big) \\ &\qquad \times\Big(\frac{1}{n}\sum_i \Pi_J(h-h_0)(X_i)\widetilde{b}^K(W_i) - \mathrm{E}[\Pi_J(h-h_0)(X)\widetilde{b}^K(W)]\Big) \\ &= T_{11} + T_{12}, \end{aligned}$$

where we used the notation $\widetilde{b}^K(\cdot) = G_b^{-1/2}b^K(\cdot)$. Consider T_{11}. We obtain

$$\begin{aligned} \mathrm{E}\,|T_{11}|^2 &\le n^{-1}\,\mathrm{E}\Big|\langle Q_J(h-h_0), \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\Pi_J(h-h_0)(X)\widetilde{b}^K(W)\Big|^2 \\ &\le 2n^{-1}\big\|\langle Q_J(h-h_0), \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\big\|^2\|\Pi_K T(h-h_0)\|_{L^2(W)}^2 \\ &\quad + 2n^{-1}\big\|\langle Q_J(h-h_0), \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\big\|^2\|\Pi_K T(h-h_0-\Pi_J(h-h_0))\|_{L^2(W)}^2 \\ &= O\Big(n^{-1}\big\|\langle Q_J(h-h_0), \psi^J\rangle_\mu'(G_b^{-1/2}S)_l^-\big\|^2\Big), \end{aligned}$$

where the second bound is due to the Cauchy-Schwarz inequality and the third bound is due to Assumption 2. To establish an upper bound for T_{12} we infer from Chen and Christensen

[2018, Lemma F.10 (c)] that

$$
\begin{aligned}
|T_{12}|^2 \le & \big\|\langle Q_J(h-h_0), \psi^J\rangle_\mu' (G_b^{-1/2}S)_l^- \big\|^2 \\
& \times \Big\| G_b^{-1/2} S' \Big((\widehat{G}_b^{-1/2}\widehat{S})_l^- \widehat{G}_b^{-1/2} G_b^{1/2} - (G_b^{-1/2}S)_l^- \Big) \Big\|^2 \\
& \times \Big\| \frac{1}{n}\sum_i b^K(W_i)\Pi_J(h-h_0)(X_i) - \mathrm{E}[\Pi_J(h-h_0)(X) b^K(W)] \Big\|^2 \\
& = \big\|\langle Q_J(h-h_0), \psi^J\rangle_\mu' (G_b^{-1/2}S)_l^- \big\|^2 \times O_p\big(n^{-1}s_J^{-2}\zeta_J^2(\log J)\big) \times O_p\big(n^{-1}\zeta_J^2\big) \\
& = O_p\Big(n^{-1}\big\|\langle Q_J(h-h_0), \psi^J\rangle_\mu' (G_b^{-1/2}S)_l^- \big\|^2\Big)
\end{aligned}
$$

using Assumption 2 (i), i.e., $s_J^{-1}\zeta_J^2\sqrt{(\log J)/n} = O(1)$. Consider the second summand on the right hand side of (F.6). Following the upper bound of T_{12} we obtain

$$
\begin{aligned}
& \Big| \mathrm{E}[(h-h_0)(X)b^K(W)]' A' G(\widehat{A}-A)\, \mathrm{E}[(h-h_0-\Pi_J(h-h_0))(X)b^K(W)] \Big|^2 \\
& \le \big\|\langle Q_J(h-h_0), \psi^J\rangle_\mu' (G_b^{-1/2}S)_l^- \big\|^2 \Big\| G_b^{-1/2} S' \Big((\widehat{G}_b^{-1/2}\widehat{S})_l^- \widehat{G}_b^{-1/2} G_b^{1/2} - (G_b^{-1/2}S)_l^- \Big) \Big\|^2 \\
& \quad \times \big\| \langle T(h-h_0-\Pi_J(h-h_0)), \widetilde{b}^K\rangle_{L^2(W)} \big\|^2 \\
& \le \big\|\langle Q_J(h-h_0), \psi^J\rangle_\mu' (G_b^{-1/2}S)_l^- \big\|^2 \|\Pi_K T(h-h_0-\Pi_J(h-h_0))\|_{L^2(W)}^2 \\
& \quad \times O_p\big(n^{-1}s_J^{-2}\zeta_J^2(\log J)\big) \\
& = O\Big(n^{-1}\big\|\langle Q_J(h-h_0), \psi^J\rangle_\mu' (G_b^{-1/2}S)_l^- \big\|^2\Big)
\end{aligned}
$$

using that $s_J^{-2}\|T(h-h_0-\Pi_J(h-h_0))\|_{L^2(W)}^2 = O(\|h-h_0-\Pi_J(h-h_0)\|_\mu^2)$ by Assumption 3 (ii) and $\zeta_J^2(\log J)\|h-\Pi_J h\|_\mu^2 = O(1)$ by Assumption 2 (ii), which implies the upper bound (F.5).

Consider T_2. We make use of the decomposition

$$
\begin{aligned}
T_2 &= \mathrm{E}[\Pi_J(h-h_0)(X)b^K(W)]'(\widehat{A}-A)'G(\widehat{A}-A)\,\mathrm{E}[\Pi_J(h-h_0)(X)b^K(W)] \\
&\quad + 2\,\mathrm{E}[\Pi_J^\perp(h-h_0)(X)b^K(W)]'(\widehat{A}-A)'G(\widehat{A}-A)\,\mathrm{E}[\Pi_J(h-h_0)(X)b^K(W)] \\
&\quad + \mathrm{E}[\Pi_J^\perp(h-h_0)(X)b^K(W)]'(\widehat{A}-A)'G(\widehat{A}-A)\,\mathrm{E}[\Pi_J^\perp(h-h_0)(X)b^K(W)] \\
&= T_{21} + T_{22} + T_{23}
\end{aligned}
$$

where we denote the projection $\Pi_J^\perp = \mathrm{id} - \Pi_J$. Consider T_{21}. We make use of the inequality

$$
\begin{aligned}
& \mathrm{E}\Big\| \Big(\frac{1}{n}\sum_i (h-h_0)(X_i)b^K(W_i) - \mathrm{E}[(h-h_0)(X)b^K(W)] \Big)' A' G^{1/2} \Big\|^2 \\
& \qquad \le n^{-1}\,\mathrm{E}\Big[(h-h_0(X))^2 \big\| b^K(W)' A' G^{1/2} \big\|^2 \Big] \\
& \qquad \le n^{-1} s_J^{-2}\sqrt{J},
\end{aligned}
$$

using the Euclidean norm is bounded by the Frobenius norm. Consequently, we get

$$\begin{aligned}
T_{21} &\leq 2\big\|G^{1/2}\big\{(\widehat{G}_b^{-1/2}\widehat{S})_l^- \widehat{G}_b^{-1/2}G_b^{1/2} - (G_b^{-1/2}S)_l^-\big\}\big\|^2 \\
&\quad \times \Big\|\frac{1}{n}\sum_i (Y_i - h_0(X_i))b^K(W_i) - \mathrm{E}[(Y - h_0(X))b^K(W)]\Big\|^2 \\
&\quad + 2\Big\|\frac{1}{n}\sum_i \big(b^K(W_i)(h-h_0)(X_i) - \mathrm{E}[(Y-h_0(X))b^K(W)]\big)' A'G^{1/2}\Big\|^2 \\
&= O_p\big(n^{-1}s_J^{-4}\zeta_J^2(\log J)\big) \times O_p\big(n^{-1}\zeta_J^2\big) + O_p\left(n^{-1}\mathrm{v}_J\right) \\
&= O_p\left(n^{-1}s_J^{-2}\sqrt{J}\right)
\end{aligned}$$

using Chen and Christensen [2018, Lemma F.10(b)] (with G_ψ replaced by G) and that $n^{-1}s_J^{-2}\zeta_J^4(\log J) = O(1)$ by Assumption 2 (i). Since $|T_{22}| \leq \sqrt{T_{21}T_{23}}$ we conclude $T_2 = O_p(n^{-1}s_J^{-2})$, which completes the proof. □

Lemma F.5. *Let Assumptions 1–5 be satisfied. Then, under $\mathcal{H}_0 = \{h_0\}$ it holds uniformly in $J \in \mathcal{I}_n$:*

$$\frac{1}{n-1}\sum_{i\neq i'}\big(Y_i - h_0(X_i)\big)\big(Y_{i'} - h_0(X_{i'})\big)b^K(W_i)'\Big(A'GA - \widehat{A}'G\widehat{A}\Big)b^K(W_{i'}) = O_p(\mathrm{v}_J).$$

Proof. We make use of the inequality

$$\begin{aligned}
&\sum_{i\neq i'}\big(Y_i - h_0(X_i)\big)\big(Y_{i'} - h_0(X_{i'})\big)b^K(W_i)'\Big(A'GA - \widehat{A}'G\widehat{A}\Big)b^K(W_{i'}) \\
&\leq \Big\|\sum_i \big(Y_i - h_0(X_i)\big)\widetilde{b}^K(W_i)\Big\|^2 \big\|G_b^{1/2}(A'GA - \widehat{A}'G\widehat{A})G_b^{1/2}\big\| \\
&= \sum_{i\neq i'}\big(Y_i - h_0(X_i)\big)\big(Y_{i'} - h_0(X_{i'})\big)\widetilde{b}^K(W_i)'\widetilde{b}^K(W_{i'})\big\|G_b^{1/2}(A'GA - \widehat{A}'G\widehat{A})G_b^{1/2}\big\| \\
&\quad + \frac{1}{n}\sum_i \Big\|\big(Y_i - h_0(X_i)\big)\widetilde{b}^K(W_i)\Big\|^2 \big\|G_b^{1/2}(A'GA - \widehat{A}'G\widehat{A})G_b^{1/2}\big\|.
\end{aligned}$$

Note that $n^{-1}\sum_i \|(Y_i - h_0(X_i))\widetilde{b}^K(W_i)\|^2 \leq K + o_p(1)$ uniformly in K. Further, by Lemma F.2 we have $\mathrm{v}_J \geq \underline{\sigma}^2\sqrt{J}$ (we may assume that $s_1 = 1$) and thus we obtain

$$\begin{aligned}
&\mathrm{P}_{h_0}\Big(\max_{J\in\mathcal{I}_n}\Big|\frac{1}{(n-1)\mathrm{v}_J}\sum_{i,i'}\big(Y_i - h_0(X_i)\big)\big(Y_{i'} - h_0(X_{i'})\big)b^K(W_i)'\Big(A'GA - \widehat{A}'G\widehat{A}\Big)b^K(W_{i'})\Big| > 1\Big) \\
&\leq \mathrm{P}_{h_0}\Big(\max_{J\in\mathcal{I}_n}\Big|\frac{1}{n\sqrt{J}}\sum_{i\neq i'}\big(Y_i - h_0(X_i)\big)\big(Y_{i'} - h_0(X_{i'})\big)\widetilde{b}^K(W_i)'\widetilde{b}^K(W_{i'})\Big| > \underline{\sigma}^{-2}\Big) \\
&\quad + \mathrm{P}_{h_0}\Big(\max_{J\in\mathcal{I}_n}\big(\|G_b^{1/2}(A'GA - \widehat{A}'G\widehat{A})G_b^{1/2}\|\big) > \underline{\sigma}^{-2}/2\Big) \\
&\quad + \mathrm{P}_{h_0}\Big(\max_{J\in\mathcal{I}_n}\big(K\|G_b^{1/2}(A'GA - \widehat{A}'G\widehat{A})G_b^{1/2}\|\big) > \underline{\sigma}^{-2}/2\Big) + o(1).
\end{aligned}$$

Consequently, the result follows from the proofs of Theorem 3.3 and Chen and Christensen [2015, Lemma E.16]. □

Lemma F.6. *Let Assumptions 1–5 be satisfied. Then, for some* $c_0 \in (0,1)$ *it holds* $|\widehat{\mathrm{v}}_J - \mathrm{v}_J| \leq c_0 \mathrm{v}_J$ *wpa1 for all* $J \in \mathcal{I}_n$.

Proof. We denote $G_{\sigma^2} = \mathrm{E}_h[(Y - h_0(X))^2 b^K(W) b^K(W)']$ and its empirical analog $\widehat{G}_{\sigma^2} = n^{-1}\sum_i (Y_i - h_0(X_i))^2 b^K(W_i) b^K(W_i)'$. Note that for any $J \times J$ matrix M it holds $\|M\|_F \leq \sqrt{J}\|M\|$ and hence For all $J \in \mathcal{I}_n$ the triangular inequality implies

$$
\begin{aligned}
|\widehat{\mathrm{v}}_J - \mathrm{v}_J| &\leq \left\| G^{1/2}\widehat{A}\widehat{G}_{\sigma^2}\widehat{A}'G^{1/2} - G^{1/2}AG_{\sigma^2}A'G^{1/2} \right\|_F \\
&\leq \sqrt{J}\left\| G^{1/2}\widehat{A}\widehat{G}_{\sigma^2}\widehat{A}'G^{1/2} - G^{1/2}AG_{\sigma^2}A'G^{1/2} \right\|.
\end{aligned}
$$

Thus, the result follows from the proof of Chen and Christensen [2015, Lemma E.16]. □

Lemma F.7. *Let Assumptions 1–5 be satisfied. Then, we have* $\max_{J \in \mathcal{I}_n} \left| 1 - \frac{\widehat{\mathrm{v}}_J(h_0)}{\widehat{\mathrm{v}}_J(\widehat{h}_J)} \right| = o_p(1)$.

Proof. For all $J \in \mathcal{I}_n$ and $h \in \mathcal{H}_J^{\mathrm{R}}$ the triangular inequality implies

$$
\begin{aligned}
&|\widehat{\mathrm{v}}_J(h) - \widehat{\mathrm{v}}_J(h_0)| \\
&\leq \left\| G^{1/2}\widehat{A}\frac{1}{n}\sum_i \big((Y_i - h(X_i))^2 - (Y_i - h_0(X_i))^2\big) b^K(W_i) b^K(W_i)'\widehat{A}'G^{1/2} \right\|_F \\
&\leq \sqrt{J}\left\| G^{1/2}\widehat{A}\frac{1}{n}\sum_i \big(h(X_i) - h_0(X_i)\big)^2 b^K(W_i) b^K(W_i)'\widehat{A}'G^{1/2} \right\| \\
&\quad + 2\sqrt{J}\left\| G^{1/2}\widehat{A}\frac{1}{n}\sum_i \big(h(X_i) - h_0(X_i)\big)(Y_i - h_0(X_i)) b^K(W_i) b^K(W_i)'\widehat{A}'G^{1/2} \right\| \\
&= T_1 + T_2.
\end{aligned}
$$

Consider T_1. Following the proof of Theorem 3.4 we obtain

$$
T_1 \leq \sqrt{J}s_J^{-2}\left\| n^{-1}\sum_i \big(h(X_i) - h_0(X_i)\big)^2 b^K(W_i) b^K(W_i)' \right\| = O_p(\sqrt{J}s_J^{-2}n^{-1}) = o_p(1)
$$

uniformly in $J \in \mathcal{I}_n$ by Assumption (5) (i). Analogously, we obtain $T_2 = o_p(1)$ uniformly in $J \in \mathcal{I}_n$. □

Lemma F.8. *Under Assumption 1 (ii) and 2 (iii) it holds for all* $h \in \mathcal{H}$ *that*

$$
\lambda_{\max}\big(\mathbb{V}ar_h(\rho(Y, h_0(X))G_b^{-1/2}b^K(W))\big) \leq \overline{\sigma}^2 < \infty.
$$

Proof. For any $\gamma \in \mathbb{R}^K$ it holds

$$\begin{aligned}\gamma' \mathbb{V}ar_h\left(\rho(Y, h_0(X)) G_b^{-1/2} b^K(W)\right)\gamma &\leq \mathrm{E}\left[\mathrm{E}_h[\rho^2(Z, h_0)|W]\left(\gamma' G_b^{-1/2} b^K(W)\right)^2\right] \\ &\leq \overline{\sigma}^2\, \mathrm{E}\left[\left(\gamma' G_b^{-1/2} b^K(W)\right)^2\right] \\ &= \overline{\sigma}^2 \gamma' G_b^{-1/2}\, \mathrm{E}\left[b^K(W) b^K(W)'\right] G_b^{-1/2}\gamma = \overline{\sigma}^2 \|\gamma\|^2,\end{aligned}$$

where the second inequality is due to to Assumption 1 (ii). □

Lemma F.9. *Under the conditions of Theorem 4.1 it holds*

$$\widehat{D}_K - \|m(\cdot, h_0)\|^2_{L^2(W)} = O_p\left(n^{-1}\sqrt{K} + n^{-1/2}\|\Pi_K m(\cdot, h_0)\|_{L^2(W)} + K^{-2\gamma/d_w}\right).$$

Proof. Similarly to Theorem F.1 we obtain

$$\widehat{D}_K - \|m(\cdot, h_0)\|^2_{L^2(W)} = \widehat{D}_K - \|\Pi_K m(\cdot, h_0)\|^2_{L^2(W)} + \|\Pi_K m(\cdot, h_0)\|^2_{L^2(W)} - \|m(\cdot, h_0)\|^2_{L^2(W)}.$$

Following the first part of the proof of Theorem F.1 with V_i^K replaced by $\rho(Y_i, h_0(X_i)) G_b^{-1/2} b^K(W_i)$ and using Lemma F.4 yields

$$\widehat{D}_K - \|\Pi_K m(\cdot, h_0)\|^2_{L^2(W)} = O\left(n^{-1}\sqrt{K} + n^{-1/2}\|\Pi_K m(h_0, \cdot)\|_{L^2(W)}\right).$$

Moreover, we obtain using that Π_K is a projection on $L^2(W)$:

$$\begin{aligned}&\|\Pi_K m(\cdot, h_0)\|^2_{L^2(W)} - \|m(\cdot, h_0)\|^2_{L^2(W)} \\ &\quad = \|\Pi_K m(\cdot, h_0) - m(\cdot, h_0)\|^2_{L^2(W)} + 2\langle \Pi_K m(\cdot, h_0) - m(\cdot, h_0), \Pi_K m(\cdot, h_0)\rangle_{L^2(W)} \\ &\quad = 3\|\Pi_K m(\cdot, h_0) - m(\cdot, h_0)\|^2_{L^2(W)} \\ &\quad = O(K^{-2\gamma/d_w}),\end{aligned}$$

where the last bound is due to the sieve approximation rate imposed in Assumption 7. □

G. U-statistics deviation results

We make use of the following exponential inequality established by Houdré and Reynaud-Bouret [2003].

Lemma G.1 (Houdré and Reynaud-Bouret [2003]). *Let U_n be a degenerate U-statistic of order 2 with kernel R based on a simple random sample $Z_1, \ldots, Z_n$. Then there exists a* *generic constant $C > 0$, such that*

$$\mathrm{P}_h\left(\Big|\sum_{1\le i<i'\le n} R(Z_i, Z_{i'})\Big| \ge C\Big(\Lambda_1\sqrt{u} + \Lambda_2 u + \Lambda_3 u^{3/2} + \Lambda_4 u^2\Big)\right) \le 6\exp(-u)$$

where

$$\begin{aligned}
\Lambda_1^2 &= \frac{n(n-1)}{2}\,\mathrm{E}[R^2(Z_1, Z_2)],\\
\Lambda_2 &= n \sup_{\|\nu\|_{L^2(Z)}\le 1, \|\kappa\|_{L^2(Z)}\le 1} \mathrm{E}[R(Z_1, Z_2)\nu(Z_1)\kappa(Z_2)],\\
\Lambda_3 &= \sqrt{n \sup_z |\,\mathrm{E}[R^2(Z_1, z)]|},\\
\Lambda_4 &= \sup_{z_1, z_2} |R(z_1, z_2)|.
\end{aligned}$$

The next result provides upper bounds for the estimates $\Lambda_1, \ldots, \Lambda_4$ when the kernel R coincides with R_1 given in Appendix B. Also from Appendix B recall the definition $Z_i = (Y_i, X_i, W_i)$ and $M_i = \{|Y_i - h_0(X_i)| \le M_n\}$. Recall that the kernel R_1 is a symmetric function satisfying $\mathrm{E}[R_1(Z, z)] = 0$ for all z.

Lemma G.2. *Let Assumption 1 (ii) be satisfied. Given kernel R_1 it holds under $\mathcal{H}_0$:*

$$\Lambda_1^2 \le \frac{n(n-1)}{2}\, v_J^2, \tag{G.1}$$

$$\Lambda_2 \le 2\overline{\sigma}^2\, n\, s_J^{-2}, \tag{G.2}$$

$$\Lambda_3 \le \overline{\sigma}^2 \sqrt{n}\, M_n\, \zeta_{b,K}\, s_J^{-2}, \tag{G.3}$$

$$\Lambda_4 \le M_n^2 \zeta_{b,K}^2 s_J^{-2}. \tag{G.4}$$

Proof. Proof of (G.1). Recall the notation $V_i^J = U_i G^{1/2} A b^K(W)$ with $U_i = Y_i - h(X_i)$, then we evaluate under $\mathcal{H}_0$:

$$\begin{aligned}
\mathrm{E}_h[R_1^2(Z_1, S_2)] &\le \mathrm{E}_h\left|U_1 b^K(W_1)' A' G A b^K(W_2) U_2\right|^2\\
&= \mathrm{E}_h\left[U^2 b^K(W)' A' G A\, \mathrm{E}_h\left[U^2 b^K(W) b^K(W)'\right] A' G A b^K(W)\right]\\
&= \mathrm{E}_h\left[(V^J)'\, \mathrm{E}_h\left[V^J (V^J)'\right] V^J\right] = \sum_{j,j'=1}^J \mathrm{E}_h[V_j V_{j'}]^2 = v_J^2.
\end{aligned}$$

Proof of (G.2). For any function ν and κ with $\|\nu\|_{L^2(Z)} \le 1$ and $\|\kappa\|_{L^2(Z)} \le 1$, respectively, we obtain

$$
\begin{aligned}
|\operatorname{E}_h[R_1(Z_1,Z_2)\nu(Z_1)\kappa(Z_2)]| \leq & \Big|\operatorname{E}[U\mathbb{1}_M b^K(W)'\nu(Z)]A'GA\operatorname{E}_h[U\mathbb{1}_M b^K(W)\kappa(Z)]\Big| \\
& + \big\|G^{1/2}A\operatorname{E}_h[U\mathbb{1}_M b^K(W)]\big\|^2 \\
\leq & \|G^{1/2}A\operatorname{E}_h[U\mathbb{1}_M b^K(W)\kappa(Z)]\|\,\|G^{1/2}A\operatorname{E}_h[U\mathbb{1}_M b^K(W)\nu(Z)]\| \\
& + \big\|G^{1/2}A\operatorname{E}_h[U\mathbb{1}_M b^K(W)]\big\|^2 \\
\leq & \|G^{1/2}AG_b^{1/2}\|^2\Big(\sqrt{\operatorname{E}\left[|\operatorname{E}_h[U\mathbb{1}_M\kappa(Z)|W]|^2\right]}\times\sqrt{\operatorname{E}_h\left[|\operatorname{E}_h[U\mathbb{1}_M\nu(Z)|W]|^2\right]} \\
& \quad + \operatorname{E}\left[|\operatorname{E}_h[U\mathbb{1}_M|W]|^2\right]\Big).
\end{aligned}
$$

Now observe $\operatorname{E}\left[|\operatorname{E}_h[U\mathbb{1}_M\kappa(Z)|W]|^2\right] \leq \operatorname{E}\left[\operatorname{E}_h[U^2|W]\kappa^2(Z)\right] \leq \overline{\sigma}^2$ by Assumption 1 (ii) and using that $\|\kappa\|_{L^2(Z)} \leq 1$, which yields the upper bound by using $\|G^{1/2}AG_b^{1/2}\| = s_J^{-1}$.

Proof of (G.3). Observe that for any $z = (u, w)$

$$
\begin{aligned}
\big|\operatorname{E}_h[R_1^2(Z_1,z)]\big| &\leq \operatorname{E}_h\Big|U\mathbb{1}\{|U|\leq M_n\}b^K(W)'A'GAb^K(w)u\mathbb{1}\{|u|\leq M_n\}\Big|^2 \\
&\leq \|G^{1/2}Ab^K(w)u\mathbb{1}\{|u|\leq M_n\}\|^2\operatorname{E}_h\|G^{1/2}Ab^K(W)U\|^2 \\
&\leq \overline{\sigma}^2M_n^2\,\zeta_{b,K}^2\,\|G^{1/2}AG_b^{1/2}\|^4,
\end{aligned}
$$

again by using Assumption 1 (ii) and hence the upper bound (G.3) follows.

Proof of (G.4). Observe that for any $z_1 = (u_1, w_1)$ and $z_2 = (u_2, w_2)$ we get

$$
\begin{aligned}
\big|R_1(s_1,s_2)\big| &\leq \Big|u_1\mathbb{1}\{|u_1|\leq M_n\}b^K(w_1)'A'GAb^K(w_2)u_2\mathbb{1}\{|u_2|\leq M_n\}\Big| \\
&\leq \sup_{u,w}\big\|G^{1/2}Ab^K(w)u\mathbb{1}\{|u|\leq M_n\}\big\|^2 \leq M_n^2\zeta_{b,K}^2\|G^{1/2}AG_b^{1/2}\|^2,
\end{aligned}
$$

which completes the proof. □

References

X. Chen and T. Christensen. Optimal sup-norm rates, adaptivity and inference in nonparametric instrumental variables estimation, 2015.

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

X. Chen and D. Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 01 2012.

C. Houdré and P. Reynaud-Bouret. Exponential inequalities, with constants, for u-statistics of order two. In *Stochastic inequalities and applications*, pages 55–69. Springer, 2003.