6-16-2020

Adaptive, Rate-Optimal Hypothesis Testing in Nonparametric IV Models

Christoph Breunig

Xiaohong Chen
Yale University

ADAPTIVE, RATE-OPTIMAL HYPOTHESIS TESTING IN NONPARAMETRIC IV MODELS

By

Christoph Breunig and Xiaohong Chen

June 2020

Revised December 2021

COWLES FOUNDATION DISCUSSION PAPER NO. 2238R

Adaptive, Rate-Optimal Hypothesis Testing in Nonparametric IV Models[*]

CHRISTOPH BREUNIG[†]   XIAOHONG CHEN[‡]

First version: August 2018, Revised December 22, 2021

We propose a new adaptive hypothesis test for polyhedral cone (e.g., monotonicity, convexity) and equality (e.g., parametric, semiparametric) restrictions on a structural function in a nonparametric instrumental variables (NPIV) model. Our test statistic is based on a modified leave-one-out sample analog of a quadratic distance between the restricted and unrestricted sieve NPIV estimators. We provide computationally simple, data-driven choices of sieve tuning parameters and adjusted chi-squared critical values. Our test adapts to the unknown smoothness of alternative functions in the presence of unknown degree of endogeneity and unknown strength of the instruments. It attains the adaptive minimax rate of testing in L^2. That is, the sum of its type I error uniformly over the composite null and its type II error uniformly over nonparametric alternative models cannot be improved by any other hypothesis test for NPIV models of unknown regularities. Data-driven confidence sets in L^2 are obtained by inverting the adaptive test. Simulations confirm that our adaptive test controls size and its finite-sample power greatly exceeds existing non-adaptive tests for monotonicity and parametric restrictions in NPIV models. Empirical applications to test for shape restrictions of differentiated products demand and of Engel curves are presented.

Keywords: Nonparametric instrumental variables; Shape restrictions; Nonparametric alternatives; Minimax rate of testing; Adaptive hypothesis testing; Random exponential scan; Sieve regularization; Quadratic functionals.

[†]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

In this paper, we propose computationally simple, data-driven optimal hypothesis testing in a nonparametric instrumental variables (NPIV) model. The maintained assumption is that there is a nonparametric structural function h satisfying the NPIV model

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

where X is a d_x-dimensional vector of possibly endogenous regressors, W is a d_w-dimensional vector of conditional (instrumental) variables (with $d_w \geq d_x$), and the joint distribution of (Y, X, W) is unspecified beyond (1.1). With the danger of abusing terminology, we call a function h satisfying model (1.1) a NPIV function. We are interested in testing a null hypothesis that a NPIV function h satisfies some simplifying economic restrictions, such as parametric or semiparametric equality restrictions or polyhedral cone restrictions (e.g., nonnegativity, monotonicity, convexity or supermodularity). Our new test builds on a simple data-driven choice of tuning parameters that ensures asymptotic size control and non-trivial power uniformly against a large class of nonparametric alternatives.

Before presenting the theoretical properties of our new test, we derive the *minimax rate of testing* in L^2, which is the smallest rate of separation in L^2 distance between the null hypothesis and the nonparametric alternatives that ensures consistent testing uniformly over the latter. We establish the minimax result in two steps: First, we derive, for all possible tests, a lower bound for the type I error uniformly over distributions satisfying the null hypothesis and the type II error uniformly over the nonparametric alternative NPIV functions separated from the null hypothesis by a rate r_n that shrinks to zero as the sample size n goes to infinity. Thus, there exists no other test that provides a better performance with respect to the sum of those errors. Second, we propose a test whose sum of the type I and the type II errors are bounded from above (by the nominal level) at the same separation rate r_n. This test is based on a modified leave-one-out sample analog of a quadratic distance between the restricted and unrestricted sieve NPIV estimators of h. The test is shown to attain the minimax rate of testing r_n when the sieve dimension is chosen optimally according to the smoothness of the nonparametric alternative functions and the degree of the ill-posedness of the NPIV model (that depends on the smoothness of the conditional density of X given W). We call this test minimax rate-optimal (with known model regularities).

In practice, the smoothness of the nonparametric alternative functions and the degree of the ill-posedness of the NPIV model are both unknown. Our new test is a data-driven version of the minimax rate-optimal test that adapts to the unknown smoothness of the nonparametric alternative NPIV functions in the presence of the unknown degree of the ill-posedness. Our data-driven test rejects the null hypothesis as soon as there is a sieve

dimension (say the smallest sieve dimension) in an estimated index set such that the corresponding normalized quadratic distance estimator exceeds one; and fails to reject the null otherwise. The normalization builds on Bonferroni corrected chi-squared critical values, where the degree of freedom is the rank of the inequality restrictions that are active in finite samples. The cardinality of the estimated index set is determined by a random exponential scan (RES) procedure that automatically takes into account the unknown degree of ill-posedness.

We show that the new data-driven test attains the minimax rate of testing for severely ill-posed NPIV models, and is up to a $\sqrt{\log \log(n)}$ multiplicative factor of the minimax rate of testing for the mildly ill-posed NPIV models. This extra $\sqrt{\log \log(n)}$ term is the price to pay for adaptivity to unknown smoothness of nonparametric alternative functions.[1] A key technical part to establish this rate optimality in L^2 testing 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 E. We show that our adaptive test controls size by deriving a tight lower bound for Bonferroni corrected chi-squared critical values. By inverting our adaptive tests we obtain L^2 confidence sets on restricted structural functions. These confidence sets are free of additional choices of tuning parameters. The adaptive minimax rate of testing determines the L^2 radius of the confidence sets.

In Monte Carlo simulations, we analyze the finite sample properties of our adaptive test for the null of monotonicity or a parametric hypothesis using various simulation designs. Our simulations reveal the following patterns: First, our adaptive test delivers adequate size control under different composite null hypotheses and for varying strengths of the instruments. Second, our adaptive test is powerful in comparison to existing tests when alternative functions are relatively simple. Moreover, the finite-sample power of our adaptive test greatly exceeds that of existing tests when alternative functions become more nonlinear. The great power gains of our adaptive test are present even for relatively small sample sizes. For example, when comparing our adaptive test to the nonadaptive bootstrap test of Fang and Seo [2021] for the null of monotonicity, their test has trivial power against certain nonlinear alternative functions, while our adaptive test remains powerful. After combining our data-driven choice of the sieve dimension with their bootstrapped critical value, the resulting "adaptive" bootstrap test has virtually the same finite-sample power as that of our simple adaptive test. This highlights the importance of our data-driven choice of the sieve dimension to ensure powerful performance uniformly against a large class of alternative NPIV functions.

Our paper is the first about adaptive, minimax rate-optimal hypothesis testing in NPIV models that allows for a large class of semiparametric equality restrictions and polyhedral

---

[1]This is needed even for adaptive minimax hypothesis testing in nonparametric regressions (without endogeneity); see Spokoiny [1996], Horowitz and Spokoiny [2001] and Guerre and Lavergne [2005].

cone restrictions. We present two empirical applications. The first is adaptive testing for connected substitutes shape restrictions in demand for differential products using market level data. The second application is adaptive testing for monotonicity, convexity and parametric forms in Engel curves using household level data.

There are many papers on testing 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], Chen and Pouzo [2015], Tao [2020], Chernozhukov et al. [2015], Zhu [2020], Fang and Seo [2021] and the references therein. All the existing papers on testing NPIV models assume that some non-random sequences of key tuning (regularization) parameters satisfy some theoretical rate conditions. Our paper makes an important contribution to this literature by providing practical, data-driven choices of key tuning parameters in testing equality restrictions and polyhedral cone restrictions in NPIV models.

Shape restrictions play a central role in economics and econometrics; see, e.g., Matzkin [1994] and Chetverikov et al. [2018] for reviews. See Horowitz and Lee [2012], Blundell et al. [2017], Chetverikov and Wilhelm [2017], Freyberger and Reeves [2019], Compiani [2021] and the references therein for nonparametric estimation by directly imposing shape restrictions. See Chetverikov [2019], Chernozhukov et al. [2015], Zhu [2020], Fang and Seo [2021] and the references therein for testing for shapes in nonparametric and semiparametric models. Our paper contributes to this literature by providing an adaptive and rate-optimal test for shape restrictions in NPIV models. Our simulation studies and real data applications indicate that our new test is not only computationally very fast, but also has very good size and power in finite samples, without the need of computationally intensive bootstrap critical values.

The remainder of the paper is organized as follows. Section 2 describes our data-driven hypothesis test. Section 3 establishes the oracle minimax optimal rate of testing. Section 4 shows that this minimax optimal rate is attained (within a $\sqrt{\log\log(n)}$ term) by our data-driven testing procedure. Section 5 presents two simulation studies and Section 6 provides two empirical illustrations. Section 7 briefly concludes. Appendices A and B contain proofs for the results in Sections 3 and 4. The online supplementary appendices contain additional materials: Appendix C presents additional simulation results. Appendix D provides additional proofs for the results in Section 4. Appendices E and F contain additional technical lemmas and their proofs.

2. Preview of the Adaptive Testing

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 test

for NPIV type models in Subsection 2.2. The theoretical justifications are postponed to Sections 3 and 4.

2.1. Null Hypotheses and Nonparametric Alternatives

Let $\mathcal{H}$ denote some class of functions that captures some unknown degree of smoothness. The function class $\mathcal{H}$ might also contain maintained hypotheses like semiparametric structures. For instance, we impose a partial linear structure in our empirical illustration on demand for differential products in Section 6.1. Throughout the paper, $\{(Y_i, X_i, W_i)\}_{i=1}^{n}$ denotes a random sample from the distribution P_h of (Y, X, W) satisfying

$$Y = h(X) + U, \quad \text{where} \quad \mathrm{E}[U|W] = 0 \quad \text{and} \quad h \in \mathcal{H}. \tag{2.1}$$

Let $\mathcal{H}^{\mathrm{R}}$ denote a subset of functions in $\mathcal{H}$ that satisfies a conjectured restriction, which are determined by either inequality or equality restrictions.

We measure deviations on restricted and unrestricted structural functions via the squared distance $\|\phi\|_\mu := \sqrt{\mathrm{E}[\phi^2(X)\mu(X)]}$ for any function ϕ. Throughout the paper, μ is a known positive measurable function that is uniformly bounded from above and below from zero on some subset of the support of X. We let $\mu \equiv 1$ for hypothesis testing on the full support of X.

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

$$\mathcal{H}_0 := \left\{ h \in \mathcal{H} : \|h - \mathcal{H}^{\mathrm{R}}\|_\mu = 0 \right\}$$

is not empty, where we use the notation $\|h - \mathcal{H}^{\mathrm{R}}\|_\mu := \inf_{\phi \in \mathcal{H}^{\mathrm{R}}} \|h - \phi\|_\mu$. We note that $\mathcal{H}_0$ under inequality restrictions forms a closed convex cone and thus, the projection to it is unique. We now provide details on constraints we can allow for and also provide examples of testable hypotheses.

Semi-/Nonparametric Inequality Restrictions. For some known, linear mapping $M : \mathcal{H} \to L^2(X) = \{\phi : \mathrm{E}[\phi^2(X)] < \infty\}$ we introduce the class of functions satisfying an inequality restriction defined by M. That is, we consider

$$\mathcal{H}^{\mathrm{R}} = \left\{ h \in \mathcal{H} : Mh \leq 0 \right\}. \tag{IR}$$

Examples of M are differential operators, i.e., $(Mh)(x) = \partial^l h(x)$ denoting the l-th partial derivative with respect to components of x. This allows for hypotheses on NPIV functions including nonnegativity, monotonicity, convexity or supermodularity restrictions. In addition, we can allow for a partial linear structure, as the following example illustrates.

Semi-/Nonparametric Equality Restrictions. For some known function F, we consider the restricted class of functions

$$\mathcal{H}^{\mathrm{R}} = \left\{ h \in \mathcal{H} : h(\cdot) = F(\cdot; \theta, g) \text{ for some } \theta \in \Theta \text{ and } g \in \mathcal{G} \right\}, \tag{ER}$$

for a finite dimensional, compact parameter space Θ and a nonparametric function class $\mathcal{G}$. Examples include hypotheses of parametric functional form captured by $F(\cdot; \theta)$, of partial linear structure, or of additive separability. The function F is allowed to be nonlinear with respect to θ but assume that F is linear with respect to g. We assume that (ER) allows for the $\sqrt{n}$–rate of estimation of the parametric components.

To analyze the power of any test against nonparametric alternatives, we require some separation in $\| \cdot \|_\mu$–distance between the null and the class of nonparametric alternatives for all $h \in \mathcal{H}$. We consider the following class of alternatives

$$\mathcal{H}_1(\delta, r_n) := \left\{ h \in \mathcal{H} : \|h - \mathcal{H}_0\|_\mu^2 \geq \delta\, r_n^2 \right\}$$

for some constant $\delta > 0$ and a separation rate of testing $r_n > 0$ that decreases to zero as the sample size n goes to infinity. We say that a test statistic $\mathtt{T}_n$ with values in $\{0, 1\}$ is consistent uniformly over $\mathcal{H}_1(\delta, r_n)$ if $\sup_{h \in \mathcal{H}_1(\delta, r_n)} \mathrm{P}_h(\mathtt{T}_n = 0) = o(1)$.

In Section 3, we establish the *minimax (separation) rate of testing r_n* in the sense of Ingster [1993]: We propose a test that 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.
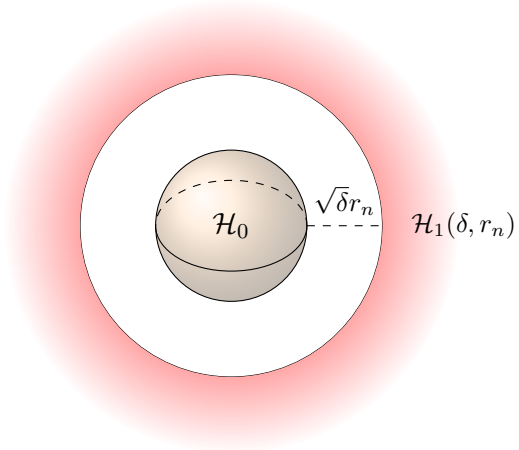
For any given level $\alpha \in (0, 1)$, a test statistic $\mathtt{T}_n$ with values in $\{0, 1\}$ is said to attain the (separation) rate of testing r_n if, for some constant $\delta^* > 0$ we have

$$\limsup_{n \to \infty} \left\{ \sup_{h \in \mathcal{H}_0} \mathrm{P}_h(\mathtt{T}_n = 1) + \sup_{h \in \mathcal{H}_1(\delta, r_n)} \mathrm{P}_h(\mathtt{T}_n = 0) \right\} \leq \alpha.$$

for all $\delta > \delta^*$. The separation rate r_n is called minimax rate of testing if for some constant

$\delta_* > 0$ it holds

$$\liminf_{n \to \infty} \inf_{T_n} \left\{ \sup_{h \in \mathcal{H}_0} P_h(T_n = 1) + \sup_{h \in \mathcal{H}_1(\delta, r_n)} P_h(T_n = 0) \right\} \geq \alpha,$$

for all $0 < \delta < \delta_*$, where $\inf_{T_n}$ is the infimum over all statistics from sample size n. Throughout the paper, we write r_n^* for the minimax rate of testing.

The minimax rate of testing r_n^* established in Section 3 depends on the unknown smoothness of $h \in \mathcal{H}$ and the inverse of the unknown conditional expectation operator T given by $Th(w) = E[h(X)|W = w]$. The minimax test in Section 3 requires an optimal choice of tuning parameters depending on these unknown objects, and hence is infeasible. In Section 4 we provide a data-driven extension to the minimax test, i.e., a testing procedure that adapts to the unknown smoothness of the unrestricted NPIV function $h \in \mathcal{H}$ in the presence of unknown smoothing properties of the conditional expectation operator T. Specifically, we propose a fully data-driven test statistic $\widehat{T}_n$ that satisfies

$$\limsup_{n \to \infty} \left\{ \sup_{h \in \mathcal{H}_0} P_h(\widehat{T}_n = 1) + \sup_{h \in \mathcal{H}_1(\delta, r_n)} P_h(\widehat{T}_n = 0) \right\} \leq \alpha, \tag{2.2}$$

where r_n is up to the $\sqrt{\log\log(n)}$ multiplicative factor of the minimax rate of testing r_n^*. We call such a feasible test $\widehat{T}_n$ *adaptive* and *rate-optimal* (or sometimes simply adaptive).

2.2. Our Adaptive Test

Our test statistic is based on a sample analog of the quadratic distance

$$D(\Pi_R h) = E\left[(h(X) - \Pi_R h(X))^2 \mu(X) \right] = \|h - \Pi_R h\|_\mu^2$$

between the NPIV function h and its projection $\Pi_R h$ onto $\mathcal{H}_0$ under the $\|\cdot\|_\mu-$norm, i.e., $\|h - \Pi_R h\|_\mu = \|h - \mathcal{H}_0\|_\mu$. More precisely, our test builds on a modified leave-one-out version of the empirical quadratic distance between the unrestricted and restricted sieve NPIV estimators of a function h satisfying model (2.1).

We first introduce sieve NPIV estimators for the NPIV function h. Let $\{\psi_j\}_{j=1}^\infty$ and $\{b_k\}_{k=1}^\infty$ be complete basis functions for spaces of square integrable functions of X and W respectively. Let $\psi^J(\cdot)$ and $b^K(\cdot)$ be vectors of basis functions of dimensions J and $K = K(J) > J$ respectively. These can be cosine, power series, spline, or wavelet basis functions. Let $\Psi = (\psi^J(X_1), \ldots, \psi^J(X_n))'$ and $B = (b^K(W_1), \ldots, b^K(W_n))'$. An unrestricted sieve NPIV estimator for the NPIV function h is given by

$$\widehat{h}_J(x) = \psi^J(x)'\widehat{\beta} \qquad \text{where} \qquad \widehat{\beta} = [\Psi'P_B\Psi]^-\Psi'P_B Y \tag{2.3}$$

where $Y = (Y_1, \ldots, Y_n)'$ and $P_B = B(B'B)^- B'$. Let $\Psi_J = \mathrm{clsp}\{\psi_1, \ldots, \psi_J\}$ and $\mathcal{H}_J^{\mathrm{R}} = \Psi_J \cap \mathcal{H}^{\mathrm{R}}$. A restricted sieve NPIV estimator is given by

$$\widehat{h}_J^{\mathrm{R}} = \arg\min_{h \in \mathcal{H}_J^{\mathrm{R}}} \sum_{i=1}^{n} \left(\widehat{h}_J(X_i) - h(X_i)\right)^2 \mu(X_i). \tag{2.4}$$

If the restricted function class is known up to a finite dimensional parameter then $\mathcal{H}_J^{\mathrm{R}} = \mathcal{H}^{\mathrm{R}}$ then the restricted estimator $\widehat{h}_J^{\mathrm{R}}$ in (2.4) does not depend on the sieve dimension J.

For each sieve dimension J, we can compute a J–dependent test statistic $n\widehat{D}_J/\widehat{v}_J$, which is a standardized, centered (or leave-one-out) version of the sample analog of $D(\Pi_{\mathrm{R}} h)$, where

$$\widehat{D}_J := \widehat{D}_J(\widehat{h}_J^{\mathrm{R}}) = \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} \left(Y - \widehat{h}_J^{\mathrm{R}}\right)'_i Q'_\Psi \Omega_\mu Q_\Psi \left(Y - \widehat{h}_J^{\mathrm{R}}\right)_{i'},$$

where $(a)_i$ coincides with a_i on the i–th entry and is zero otherwise for any $a \in \mathbb{R}^n$, $\widehat{h}_J^{\mathrm{R}} = (\widehat{h}_J^{\mathrm{R}}(X_1), \ldots, \widehat{h}_J^{\mathrm{R}}(X_n))'$, $Q_\Psi = \sqrt{n}\Psi[\Psi' P_B \Psi]^- \Psi' P_B$, and $\Omega_\mu = \mathrm{DIAG}(\mu(X_1), \ldots, \mu(X_n))$, which coincides with the identity matrix when $\mu \equiv 1$. The estimated normalization term is given by

$$\widehat{v}_J = \left\| (\Psi' \Omega_\mu \Psi)^{1/2} [\Psi' P_B \Psi]^- \Psi' P_B \, \mathrm{DIAG}(Y - \widehat{h}_J)^2 P_B \Psi [\Psi' P_B \Psi]^- (\Psi' \Omega_\mu \Psi)^{1/2} \right\|_F \tag{2.5}$$

where $\widehat{h}_J = (\widehat{h}_J(X_1), \ldots, \widehat{h}_J(X_n))'$ is the vector of the unrestricted sieve NPIV estimators[2] and $\| \cdot \|_F$ denotes the Frobenius norm.[3]

We compute our adaptive test for the null hypothesis $\mathcal{H}_0$ against nonparametric alternatives in three simple steps.

Step 1. Compute the *random exponential scan* (RES) index set:

$$\widehat{\mathcal{I}}_n = \left\{ J \le \widehat{J}_{\max} : \quad J = \underline{J} 2^j \text{ where } j = 0, 1, \ldots, j_{\max} \right\} \tag{2.6}$$

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

$$\widehat{J}_{\max} = \min \left\{ J > \underline{J} : \quad 1.5\, \zeta_J^2 \sqrt{(\log J)/n} \ge \widehat{s}_J \right\}, \tag{2.7}$$

where $\widehat{s}_J$ is the minimal singular value of $(B'B)^{-1/2}(B'\Psi)(\Psi'\Omega_\mu\Psi)^{-1/2}$. Further $\zeta_J = \sqrt{J}$ for spline, wavelet, or trigonometric sieve basis, and $\zeta_J = J$ for power series.

---

[2] We thank an anonymous referee for suggesting to use the unrestricted sieve NPIV estimator $\widehat{h}_J$ in computing $\widehat{v}_J$.

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

Step 2. Use Bonferroni correction to a critical value from a centralized chi-square distribution relative to the cardinality of the RES index set, denoted by $\#(\widehat{\mathcal{I}}_n)$, and distinguish between inequality and equality restrictions (as specified in (IR) and (ER)):

(IR) Let Ψ_{act} be a submatrix of Ψ such that $(M\Psi_{act})'\widehat{\beta}^{\mathrm{R}} = 0$, where $\widehat{\beta}^{\mathrm{R}}$ given in (2.4). Set $\widehat{\gamma}_J = \max\left(1, \mathrm{rank}(M\Psi_{act})\right)$ and compute for a given nominal level $\alpha \in (0, 1)$:

$$\widehat{\eta}_J(\alpha) = \frac{q\big(\alpha/\#(\widehat{\mathcal{I}}_n), \widehat{\gamma}_J\big) - \widehat{\gamma}_J}{\sqrt{\widehat{\gamma}_J}}, \tag{2.8}$$

where $q(a, J)$ denotes the $100(1 - a)\%$-quantile of the chi-square distribution with J degrees of freedom. Thus, the degrees of freedom of the chi-square distribution are determined by the rank of inequalities active in finite samples.

(ER) Compute $\widehat{\eta}_J(\alpha) = \big(q\big(\alpha/\#(\widehat{\mathcal{I}}_n), J\big) - J\big)/\sqrt{J}$, which corresponds to (2.8) when all constraints are binding.

Step 3. Compute $\widehat{\mathcal{W}}_J(\alpha) := \frac{n\widehat{D}_J}{\widehat{\eta}_J(\alpha)\,\widehat{v}_J}$ for all $J \in \widehat{\mathcal{I}}_n$. Compute the test

$$\widehat{\mathrm{T}}_n = \mathbb{1}\left\{\text{there exists } J \in \widehat{\mathcal{I}}_n \text{ such that } \widehat{\mathcal{W}}_J(\alpha) > 1\right\} \tag{2.9}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. Under the nominal level $\alpha \in (0, 1)$, $\widehat{\mathrm{T}}_n = 1$ indicates rejection of the null hypothesis and $\widehat{\mathrm{T}}_n = 0$ indicates a failure to reject the null.

Remark 2.1. *The RES index set $\widehat{\mathcal{I}}_n$ in Step 1 determines a collection of candidate sieve dimensions J for our test. The data-dependent upper bound $\widehat{J}_{\max}$ ensures that the cardinality of the index set $\widehat{\mathcal{I}}_n$ is not too large relative to the sampling variability of unrestricted sieve NPIV estimation. We also show that the empirical upper bound $\widehat{J}_{\max}$ diverges in probability at a rate much faster than that of $\underline{J}$ and thus, the search range is large enough to detect a large collection of alternative NPIV functions.*

Remark 2.2 (Choice of K). *We let $K = K(J) = cJ$ for some finite constant $c > 1$, and our adaptive testing procedure optimizes over J given the choice of $K(J)$. We have tried $K(J) = 2J$ and $K(J) = 4J$ in simulations. The simulation results, in terms of size and power, are not sensitive to these choices of K. This is consistent with our theory that the choice of J is the key tuning parameter in minimax rate-optimal hypothesis testing in NPIV models using sieve methods. Since our first submitted version uses $K(J) = 2J$ in simulations and empirical applications, we present empirical applications of our adaptive test using $K(J) = 4J$ in Section 6.*

3. The Minimax Rate of Testing

This section derives the minimax rate of testing in NPIV models, when $\mathcal{H}$ coincides with the Sobolev ellipsoid of *a priori* known smoothness $p > 0$. Specifically, we assume below $\mathcal{H} = \{h \in B_{2,2}^p : \|h\|_{B_{2,2}^p} \leq L\}$ for some finite radius $L > 0$, where $B_{2,2}^p$ denotes the Sobolev space with smoothness p and associated Sobolev norm $\|\cdot\|_{B_{2,2}^p}$ (see Triebel [2006, Section 1.11]). Subsection 3.1 establishes the lower bound for the rate of testing in L^2. Subsection 3.2 shows that the lower bound can be achieved by a simple test statistic if the tuning parameter can be chosen optimally.

3.1. The Lower Bound

Before we state the lower bound for the rate of testing, 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. Let $\|\phi\|_\mu^2 := \mathrm{E}[\phi^2(X)\mu(X)]$ for all $\phi \in L_\mu^2 := \{\phi : \|\phi\|_\mu < \infty\}$ associated with inner product $\langle \cdot, \cdot \rangle_\mu$. Let $T : L^2(X) \mapsto L^2(W)$ denote the conditional expectation operator given by $Th(w) = \mathrm{E}[h(X)|W = w]$.

Assumption 1. *(i)* $\inf_{w \in \mathcal{W}} \inf_{h \in \mathcal{H}} \mathbb{V}ar_h(Y - \Pi_\mathrm{R} h(X)|W = w) \geq \underline{\sigma}^2 > 0$; *(ii) for any* $h \in \mathcal{H}$, $Th = 0$ *implies that* $\|h\|_\mu = 0$; *(iii) the density of* X *is uniformly bounded below from zero and from above on* $\{x : \mu(x) > 0\}$; *and (iv) there exists a positive decreasing function* ν *such that* $\|Th\|_{L^2(W)}^2 \lesssim \sum_{j,k} [\nu(2^j)]^2 \langle h, \widetilde{\psi}_{j,k} \rangle_\mu^2$ *for all* $h \in \mathcal{H}$, *where* $\widetilde{\psi}_{j,k}$ *denotes a CDV wavelet basis.*[4]

Assumption 1(ii) is required for the 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. Assumption 1(iv) specifies the smoothing properties of the conditional expectation operator relative to the basis functions $\widetilde{\psi}_{j,k}$. This assumption is commonly imposed in the related literature, see also Chen and Reiß [2011] for an overview. As in the related literature, we distinguish between a *mildly ill-posed* case where $\nu(t) = t^{-a}$ and a *severely ill-posed* case where $\nu(t) = \exp(-t^a/2)$ for some $a > 0$. For the construction of CDV wavelet bases, we refer to Chen and Christensen [2018, Appendix E].

Theorem 3.1. *Let Assumption 1 be satisfied. Consider testing the composite hypothesis*

$$\mathcal{H}_0 = \left\{ h \in \mathcal{H} : \|h - \mathcal{H}^\mathrm{R}\|_\mu = 0 \right\} \quad versus \quad \mathcal{H}_1(\delta, r_n^*) = \left\{ h \in \mathcal{H} : \|h - \mathcal{H}_0\|_\mu \geq \sqrt{\delta}\, r_n^* \right\}.$$

---

[4]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 $b_n \lesssim a_n$.

Then for any $\alpha > 0$ there exists a constant $\delta_* > 0$ such that

$$\liminf_{n \to \infty} \inf_{\mathrm{T}_n} \left\{ \sup_{h \in \mathcal{H}_0} \mathrm{P}_h(\mathrm{T}_n = 1) + \sup_{h \in \mathcal{H}_1(\delta, r_n^*)} \mathrm{P}_h(\mathrm{T}_n = 0) \right\} \geq \alpha,$$

for all $0 < \delta < \delta_*$, where r_n^* is given by:

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

2. *Severely ill-posed case: $r_n^* = (\log n)^{-p/a}$*

and where $\sup_{h \in \mathcal{H}} \mathrm{P}_h(\cdot)$ denotes the supreme over $h \in \mathcal{H}$ and distributions of (X, W, U) satisfying Assumption 1.

Remark 3.1. *Theorem 3.1 implies that the L^2–rate of testing $(\log n)^{-p/a}$ in the severely ill-posed case coincides with the lower bound of the L^2–rate of estimation (Chen and Reiß [2011]). For the mildly ill-posed NPIV models, the L^2–rate of testing $r_n^* = n^{-2p/(4(p+a)+d_x)}$ goes to zero faster than $n^{-p/(2(p+a)+d_x)}$, which is the lower bound of the L^2–rate of estimation (Hall and Horowitz [2005] and Chen and Reiß [2011]). Therefore, minimax L^2–testing and minimax L^2–estimation are equally difficult in the severely ill-posed case, but minimax L^2–testing is easier than minimax L^2–estimation in the mildly ill-posed case.*

The lower bound in Theorem 3.1 does not require restrictions on the class of null functions $\mathcal{H}_0$. However, one needs some restrictions on the complexity of $\mathcal{H}_0$ to establish upper bounds for the rate of testing. We shall derive an upper bound under simple null hypotheses in the next subsection, and consider composite testing problems in Section 4.

3.2. An Upper Bound under Simple Hypotheses

We first consider the simple hypothesis case where $\mathcal{H}_0 = \{h_0\}$ for some known function h_0 satisfying (1.1). We introduce a J dependent analog to the adaptive test $\widehat{\mathrm{T}}_n$ under the simple null:

$$\mathrm{T}_{n,J} = \mathbb{1} \left\{ \frac{n \widehat{D}_J(h_0)}{\widehat{v}_J} > \eta_J(\alpha) \right\} \tag{3.1}$$

where $\eta_J(\alpha) = (q(\alpha, J) - J)/\sqrt{J}$ and $q(a, J)$ denotes the upper a-quantile of the chi-square distribution with J degrees of freedom. The test $\mathrm{T}_{n,J}$ with optimally chosen J serves as a benchmark of our adaptive testing procedure (given in (4.1)) under simple hypotheses.

We introduce additional notation. Let $G = \mathrm{E}[\psi^J(X)\psi^J(X)'\mu(X)]$, $G_b = \mathrm{E}[b^K(W)b^K(W)']$ and $S = \mathrm{E}[b^K(W)\psi^J(X)']$. We assume throughout that G, G_b and S have full rank. Let s_J denote the minimal singular value of $G_b^{-1/2} S G^{-1/2}$. Let $\zeta_J = \max(\zeta_{\psi,J}, \zeta_{b,K})$, $\zeta_{\psi,J} = \sup_x \|G^{-1/2}\psi^J(x)\|$ and $\zeta_{b,K} = \sup_w \|G_b^{-1/2} b^K(w)\|$, where $\|\cdot\|$ denotes the Euclidean

norm when applied to vectors and the operator norm induced by the Euclidean norm when applied to matrices. Finally we introduce the projections $\Pi_J h(\cdot) = \psi^J(\cdot)' G^{-1} \langle \psi^J, h \rangle_\mu$ for $h \in L_\mu^2$ and $\Pi_K m(\cdot) = b^K(\cdot)' G_b^{-1} E[b^K(W) m(W)]$ for $m \in L^2(W)$.

Assumption 2. *(i)* $\sup_{w \in \mathcal{W}} \sup_{h \in \mathcal{H}} E[(Y - h(X))^2 | W = w] \leq \bar{\sigma}^2 < \infty$ *and* $\sup_{h \in \mathcal{H}} E[(Y - h(X))^4] < \infty$; *(ii)* $s_J^{-1} \zeta_J^2 \sqrt{(\log J)/n} = O(1)$; *(iii)* $\zeta_J \sqrt{\log J} = O(J^{p/d_x})$; *(iv)* $\|T(h - \Pi_R h - \Pi_J(h - \Pi_R h))\|_{L^2(W)} = O(s_J \|h - \Pi_R h - \Pi_J(h - \Pi_R h)\|_\mu)$ *for all NPIV functions* $h \in \mathcal{H}$; *and (v)* ψ^J *is a sieve basis such that for some constant* $C > 0$: $\sup_{h \in \mathcal{H}} \|\Pi_J h - h\|_\mu \leq C J^{-p/d_x}$ *for all J, and* $\zeta_J = O(\sqrt{J})$ *or* $\zeta_J = O(J)$.

Assumption 2(i) captures second moment bounds. Assumption 2(ii) imposes bounds on the growth of the basis functions relative to the singular values of the matrix $G_b^{-1/2} S G^{-1/2}$. Assumption 2(ii)(iii) imposes bounds on the growth of the basis functions which are known for commonly used bases. Assumption 2(v) is satisfied by cosine, spline, wavelet basis or power series. 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] and Huang [1998]. Assumption 2(iv) is the usual L^2 "stability condition" imposed in the NPIV literature when $\Pi_R h = 0$, and is automatically satisfied by Riesz bases (cf. Blundell et al. [2007, Assumption 6] and Chen and Pouzo [2012, Assumption 5.2(ii)]).

Theorem 3.2. *Let Assumptions 1(i)-(iii) and 2 be satisfied. Consider testing the simple hypothesis*

$$\mathcal{H}_0 = \{h_0\} \quad versus \quad \mathcal{H}_1(\delta, r_{n,J}) = \Big\{ h \in \mathcal{H} : \|h - h_0\|_\mu^2 \geq \delta \, r_{n,J}^2 \Big\},$$

for a known function h_0. *Then, for any* $\alpha \in (0,1)$ *there is a constant* $\delta^* > 0$ *such that*

$$\limsup_{n \to \infty} \Big\{ P_{h_0}(T_{n,J} = 1) + \sup_{h \in \mathcal{H}_1(\delta, r_{n,J})} P_h(T_{n,J} = 0) \Big\} \leq \alpha, \tag{3.2}$$

for all $\delta > \delta^*$, *where the separation rate* $r_{n,J}$ *is given by*

$$r_{n,J} = n^{-1/2} s_J^{-1} J^{1/4} + J^{-p/d_x}. \tag{3.3}$$

Theorem 3.2 shows that the test statistic $T_{n,J}$ given in (3.1) attains the L^2- rate of testing $r_{n,J}$. Given a sieve dimension J, this rate consists of a standard deviation term $(n^{-1/2} s_J^{-1} J^{1/4})$ and a bias term (J^{-p/d_x}). 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 E.1 in the online appendix, which we show is sufficient for the consistency of $T_{n,J}$ uniformly over $\mathcal{H}_1(\delta, r_{n,J})$.

Remark 3.2 (Relation to the L^2–Rate of Sieve Estimation). *Given a sieve dimension J, the L^2–rate of sieve estimation for any NPIV function $h \in \mathcal{H}$ is given by: $n^{-1/2}s_J^{-1}J^{1/2}+J^{-p/d_x}$. Compared the L^2-rate of estimation and of testing via the sieve NPIV procedures, while both have the same bias term J^{-p/d_x}, the L^2 rate of testing has a smaller "standard deviation" term $n^{-1/2}s_J^{-1}J^{1/4}$. Intuitively, we may obtain a higher precision in testing as the L^2 rate of testing is determined by estimation of a quadratic norm of the unrestricted NPIV function $h \in \mathcal{H}$. Although this leads to a faster optimal L^2 rate of sieve testing for mildly illposed NPIV models, the optimal L^2 rate of sieve testing is the same as the optimal L^2 rate of sieve estimation for severely illposed NPIV models; see Corollary 3.1 below.*

Below, we make use of sieve L^2 measure of ill-posedness which 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: $\tau_j \sim j^{a/d_x}$ for some $a > 0$ and severely ill-posed if: $\tau_j \sim \exp(j^{a/d_x}/2)$ for some $a > 0$. We further define $\Psi_J = \mathrm{clsp}\{\psi_1, \ldots, \psi_J\} \subset L^2(X)$.

Assumption 3. $\sup_{h \in \Psi_J} \tau_J \|(\Pi_K T - T)h\|_{L^2(W)}/\|h\|_\mu = o(1)$.

Assumption 3 is a mild condition on the approximation properties of the basis used for the instrument space, see Chen and Christensen [2018, Assumption 4(i)]. Assumption 3 implies that $\tau_J \geq const. \times s_J^{-1}$. The next corollary provides concrete rates of testing in L^2 when the sieve dimension parameter J is chosen optimally to balance the variance and square bias. The resulting sieve dimension choice satisfies $J_* = \max\{J : n^{-1}s_J^{-2}\sqrt{J} \leq J^{-2p/d_x}\}$.

Corollary 3.1. *Let Assumptions 1(i)-(iii), 2, and 3 be satisfied. Then the rate of testing r_{n,J_*} given in (3.3) is of the following form:*

1. *Mildly ill-posed case: $J_* \sim n^{2d_x/(4(p+a)+d_x)}$ implies $r_{n,J_*} = r_n^* = n^{-2p/(4(p+a)+d_x)}$,*

2. *Severely ill-posed case: $J_* = (c \log n)^{d_x/a}$ for some $c \in (0,1)$ implies $r_{n,J_*} = r_n^* = (\log n)^{-p/a}$.*

Corollary 3.1 shows that the sieve test T_{n,J_*} achieves the L^2- *minimax rate of testing* for a simple null hypothesis, assuming known smoothness p of the nonparametric alternatives.

4. Adaptive Inference

This section presents several results on data-driven test statistics. We see that our test is able to adapt to the unknown smoothness $p > 0$ of the Sobolev ellipsoid $\mathcal{H}$. Subsection 4.1

establishes the rate optimality of our adaptive test for simple null hypotheses. Subsection 4.2 extends this result to testing for composite null problems. Subsection 4.3 proposes data-driven confidence sets by inverting the adaptive test under imposed restrictions on the NPIV function.

4.1. Adaptive Testing for Simple Hypotheses

We establish an upper bound for the rate of testing using our data-driven test statistic for a simple null. Under the simple null hypothesis $\mathcal{H}_0 = \{h_0\}$, for some known function h_0 satisfying (1.1), our data-driven test given in (2.9) simplifies to

$$\widehat{\mathrm{T}}_n = \mathbb{1}\left\{ \text{there exists } J \in \widehat{\mathcal{I}}_n \text{ such that } \frac{n\widehat{D}_J(h_0)}{\widehat{v}_J} > \widehat{\eta}_J(\alpha) \right\}, \tag{4.1}$$

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

Recall the definition of the RES index set $\widehat{\mathcal{I}}_n$ given in (2.6), which relies on an upper bound $\widehat{J}_{\max}$ given in (2.7). To establish our asymptotic results below, we introduce a non-random index set $\mathcal{I}_n$ with a deterministic upper bound $\overline{J}$ as follows:

$$\mathcal{I}_n = \left\{ J \leq \overline{J}: \quad J = \underline{J}2^j \text{ where } j = 0, 1, \ldots, j_{\max} \right\} \subset [\underline{J}, \overline{J}], \tag{4.2}$$

with $\overline{J} = \sup\{J : \zeta_J^2 \sqrt{(\log J)/n} \leq \overline{c}\, s_J\}$ for some sufficiently large constant $\overline{c} > 0$. We show in Lemma E.10(i) that $\widehat{J}_{\max} \leq \overline{J}$ (and thus $\widehat{\mathcal{I}}_n \subset \mathcal{I}_n$) holds with probability approaching one uniformly over all functions $h \in \mathcal{H}$. Thus $\overline{J}$ serves as a deterministic upper bound for the RES index set $\widehat{\mathcal{I}}_n$. Finally we note that $\widehat{v}_J$ given in (2.5) estimates the population normalization factor:

$$v_J = \left\| G^{1/2} [S'G_b^{-1}S]^{-1} S'G_b^{-1}\Sigma G_b^{-1}S[S'G_b^{-1}S]^{-1}G^{1/2} \right\|_F,$$

where $\Sigma = \mathrm{E}_h\big[\big(Y - h(X)\big)^2 b^K(W)b^K(W)'\big]$.

Assumption 4. *(i) Assumptions 2(ii)(iv) hold uniformly for all $J \in \mathcal{I}_n$; (ii) For all $J = J(n)$ and $L = L(n)$ with $L = o(J)$ and $L \to \infty$ it holds that $\max\big(v_L, s_L^{-2}\sqrt{\log\log L}\big) = o(v_J)$; (iii) $p \geq 3d_x/4$ when using cosine, spline, or wavelet basis functions and $p \geq 7d_x/4$ when using power series basis functions.*

Assumptions 4(i)(iii) strengthen Assumptions 2(ii)(iii)(iv) to hold uniformly over the deterministic index set. In particular, it restricts the growth of the deterministic upper bound $\overline{J}$ of the RES index set $\widehat{\mathcal{I}}_n$. Assumption 4(iii) imposes a lower bound on the smoothness of the function class $\mathcal{H}$. Assumption 4(ii) is automatically satisfied in the mildly or severely ill-posed case as we show below in the proof of Corollary 4.1.

Let an integer J_0 be the largest sieve dimension parameter such that the squared bias dominates the variance within a $\sqrt{\log\log n}$ term, that is,

$$J_0 = \max\left\{ J: \ n^{-1}\sqrt{\log\log n}\, s_J^{-2}\sqrt{J} \leq J^{-2p/d_x} \right\}. \tag{4.3}$$

Under Assumptions 4(i)(iii), Lemma E.10(ii) in the online Appendix establishes that the "optimal" adaptive sieve dimension $J_0 \in \widehat{\mathcal{I}}_n$ with probability approaching one.

Theorem 4.1. *Let Assumptions 1(i)-(iii), 2(i)(v) and 4 be satisfied. Consider testing the simple hypothesis*

$$\mathcal{H}_0 = \{h_0\} \quad versus \quad \mathcal{H}_1(\delta, r_n) = \left\{ h \in \mathcal{H}: \ \|h - h_0\|_\mu^2 \geq \delta\, r_n^2 \right\},$$

for a known function h_0. Then, for any $\alpha \in (0,1)$ there is a constant $\delta^\circ > 0$ such that

$$\limsup_{n\to\infty} \left\{ \mathrm{P}_{h_0}(\widehat{\mathsf{T}}_n = 1) + \sup_{h\in\mathcal{H}_1(\delta^\circ, r_n)} \mathrm{P}_h(\widehat{\mathsf{T}}_n = 0) \right\} \leq \alpha, \tag{4.4}$$

where the adaptive separation rate r_n is given by

$$r_n = J_0^{-p/d_x}. \tag{4.5}$$

Theorem 4.1 establishes an upper bound for the testing rate of the adaptive test $\widehat{\mathsf{T}}_n$ under a simple hypothesis. The proof of Theorem 4.1 relies on a novel exponential bound for degenerate U-statistics based on sieve estimators. In particular, we control the type I error using tight lower bounds for adjusted chi-square critical values (see Lemma E.9 in Appendix E) and show consistency of $\widehat{\mathsf{T}}_n$ uniformly over $\mathcal{H}_1(\delta^\circ, r_n)$.

We next illustrate the upper bound under classical smoothness assumptions. Again, we distinguish between the mildly or severely ill-posed case.

Corollary 4.1. *Let Assumptions 1(i)-(iii), 2(i)(v), 3 uniformly for $J \in \mathcal{I}_n$, and 4(i)(iii) be satisfied. Then, the adaptive rate of testing r_n given in (4.5) satisfies:*

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

2. *Severely ill-posed case: $r_n = (\log n)^{-p/a}$.*

From Corollary 4.1 we see that the adaptive test attains the oracle minimax rate of testing within a $\sqrt{\log\log(n)}$−term in the mildly ill-posed case. For the adaptive testing in regression models without endogeneity (i.e., when $X = W$), it is well known that the extra $\sqrt{\log\log(n)}$−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.

4.2. Adaptive Testing for Composite Hypotheses

We extend the results from the previous subsection to the case of a composite null hypothesis and, in particular, allow for testing inequality restrictions. To do so, we need to impose conditions on the complexity of the class of restricted functions $\mathcal{H}_0$.

Assumption 5. *(i) For any $\varepsilon > 0$ it holds $\sup_{h \in \mathcal{H}_0} P_h\big(\max_{J \in \mathcal{I}_n} (\zeta_J \|\widehat{h}_J^{\mathrm{R}} - h\|_\mu / c_J) > \varepsilon \big) \to 0$ with $c_J = \max\{1, (\log\log J)^{1/4}\}$; (ii) there exist a constant $C > 0$ and $J^\bullet \in \mathcal{I}_n$ with $J^\bullet \sim (J_0)^\kappa$ for some constant $0 < \kappa \leq 1$ such that $\sup_{h \in \mathcal{H}_1(\delta^\circ, r_n)} P_h\big(\|\widehat{h}_{J^\bullet}^{\mathrm{R}} - \Pi_{\mathrm{R}} h\|_\mu > C r_n\big) \to 0$, where r_n is given in (4.5); and (iii) there exists a constant $0 < c \leq 1$ such that $\inf_{h \in \mathcal{H}} P_h\big(\forall J \in \mathcal{I}_n : J^c \leq \widehat{\gamma}_J \big) \to 1$.*

Assumption 5(i)(ii) bounds the complexity of the composite null hypothesis $\mathcal{H}_0$. Assumption 5(i) is a very mild condition on the L^2–estimation rate of the restricted sieve NPIV estimator under the null hypothesis; see Remark 4.1 below for low level sufficient conditions. Assumption 5(i) implies that $\widehat{\mathsf{T}}_n$ has size control uniformly over the composite null $\mathcal{H}_0$. Assumption 5(ii) imposes that the L^2–estimation rate of the restricted sieve NPIV estimator under the alternative $\mathcal{H}_1(\delta^\circ, r_n)$ is weakly dominated by the adaptive rate of testing; see Remark 4.2 below for low level sufficient conditions. Assumption 5(ii) ensures the consistency of $\widehat{\mathsf{T}}_n$ uniformly over $\mathcal{H}_1(\delta^\circ, r_n)$. Note that Assumptions 5(i)(ii) impose estimation rate conditions on $\widehat{h}_J^{\mathrm{R}}$ under the composite null and the nonparametric alternatives, which can be viewed as NPIV extensions of the parametric estimation rate conditions imposed in Horowitz and Spokoiny [2001, Assumption 2] for testing for a parametric regression against nonparametric regressions. Assumption 5(iii) assumes that the rank of inequalities active in finite samples increases slowly with J. This assumption is automatically satisfied when all constraints are binding with $c = 1$.

Remark 4.1 (Primitive Conditions for Assumption 5(i))**.**
(1) In the case of parametric restrictions, where $\|\widehat{h}_J^{\mathrm{R}} - h\|_\mu \leq const. \times n^{-1/2}$ with probability approaching one uniformly over $\mathcal{H}_0$, Assumption 5(i) is automatically satisfied by Assumption 4(i).
(2) Under nonparametric restrictions, note that $\|\widehat{h}_J^{\mathrm{R}} - h\|_\mu$ is bounded by $\|\widehat{h}_J - h\|_\mu$ for all $h \in \mathcal{H}_0$. Assumption 5(i) implies that the RES index set has sufficient information to estimate the restricted functions $h \in \mathcal{H}_0$. Note that

$$\max_{J \in \mathcal{I}_n} \frac{\zeta_J \|\widehat{h}_J - h\|_\mu}{c_J} \leq const. \times \max_{J \in \mathcal{I}_n} \left\{ \frac{\zeta_J \sqrt{J}}{\sqrt{n} s_J c_J} + \frac{\zeta_J \|\Pi_J^{\mathcal{I}_n} h - h\|_\mu}{c_J} \right\} \tag{4.6}$$

with probability approaching one uniformly for $h \in \mathcal{H}_0$, where $\Pi_J^{\mathcal{I}_n}$ denotes the projection onto $clsp\{\psi_J : J \in \mathcal{I}_n\}$. The first summand on the right hand side of (4.6) converges to zero by the definition of $\overline{J} = \overline{J}(n)$. For the bias part, we assume that the index set has sufficient

information to approximate the NPIV function h. Let p_0 denote the smoothness and d_0 the dimension of the nonparametric component under $\mathcal{H}_0$. If $\|\Pi_J^{\mathcal{I}_n} h - h\|_\mu = O(J^{-p_0/d_0})$ and $\zeta_J = O(\sqrt{J})$, the second summand of the right hand side of (4.6) uniformly converges to zero if $p_0 \geq d_0/2$. Since the class of restricted functions $\mathcal{H}_0$ is a less complex subset of $\mathcal{H}$, it is reasonable to assume that $p_0/d_0 \geq p/d_x$ and thus $p_0 \geq d_0/2$ is automatically satisfied given Assumption 4(iii).

Remark 4.2 (Primitive Conditions for Assumption 5(ii))**.** *Assumption 5(ii) restricts the complexity of the null hypothesis to have no effect on the adaptive minimax rate of testing asymptotically, which is automatically satisfied in the severely ill-posed case (as long as the smoothness p_0 of $\mathcal{H}_0$ satisfies $p_0 \geq p$). We consider the mildly ill-posed case below. Let p_0 be the smoothness and d_0 the dimension of the nonparametric components in the class of null functions $\mathcal{H}_0$. The L^2-convergence rate of estimation is given by $n^{-1/2} J^{a/d_0 + 1/2} + J^{-p_0/d_0}$, see Remark 3.2. Choosing $J^\bullet \sim n^{d_0/(2(p_0+a)+d_0)}$ to level the variance and squared bias, the resulting L^2−rate of estimation is bounded by the optimal rate of testing $\left(\sqrt{\log \log n}/n\right)^{2p/(4(p+a)+d_x)}$ if*

$$p_0 \geq p \, \frac{4a + 2d_0}{4a + d_x}. \tag{4.7}$$

This imposes additional smoothness restrictions on $\mathcal{H}_0$. In the absence of ill-posedness, i.e., $a = 0$, and when $d_x = d_0$, the condition (4.7) is equivalent to $p_0 \geq 2p$. Larger degrees of ill-posedness relax condition (4.7). For instance, when $d_x = d_0 = 1$, $a = 1$ implies a smoothness restriction of $p_0 \geq 1.2p$ and becomes even less restrictive for larger values of a. In addition, under $d_0 \leq d_x/2$, condition (4.7) only requires $p_0 \geq p_x$. In this sense, we impose mild additional smoothness assumptions for the class of restricted functions. This is satisfied if the projection on $\mathcal{H}^{\mathrm{R}}$ leads to a modest increase of smoothness of alternative functions $h \in \mathcal{H}_1(\delta^\circ, r_n)$.

The next result establishes an upper bound for the rate of testing for a composite null hypothesis using the test statistic $\widehat{\mathrm{T}}_n$ given in (2.9).

Theorem 4.2. *Let Assumptions 1(i)-(iii), 2(i)(v), 4, and 5 be satisfied. Consider testing the composite hypothesis*

$$\mathcal{H}_0 = \left\{ h \in \mathcal{H} : \|h - \mathcal{H}^{\mathrm{R}}\|_\mu = 0 \right\} \quad \text{versus} \quad \mathcal{H}_1(\delta, r_n) = \left\{ h \in \mathcal{H} : \|h - \mathcal{H}_0\|_\mu^2 \geq \delta \, r_n^2 \right\}.$$

Then, for any $\alpha \in (0, 1)$ there exists a constant $\delta^\circ > 0$ such that

$$\limsup_{n \to \infty} \left\{ \sup_{h \in \mathcal{H}_0} \mathrm{P}_h\big(\widehat{\mathrm{T}}_n = 1\big) + \sup_{h \in \mathcal{H}_1(\delta^\circ, r_n)} \mathrm{P}_h\big(\widehat{\mathrm{T}}_n = 0\big) \right\} \leq \alpha, \tag{4.8}$$

where the adaptive rate r_n is given in Theorem 4.1. If in addition Assumption 3 holds uniformly for $J \in \mathcal{I}_n$, then

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

2. *Severely ill-posed case:* $r_n = (\log n)^{-p/a}$.

Theorem 4.2 states that $\widehat{\mathsf{T}}_n$ attains the same adaptive rate of testing r_n for a composite null as that for a simple null.

Adaptive Testing in Semiparametric Models. Partially parametric models are often used in empirical work and can be easily incorporated in our framework either as restricted models or as the maintained models. Let $\Theta \oplus \mathcal{G} = \{h(x_1, x_2) = x_1'\theta + g(x_2) : \theta \in \Theta, g \in \mathcal{G}\}$, where Θ denotes a finite dimensional parameter space, and $\mathcal{G}$ denotes a class of nonparametric functions (say $B_{2,2}^{p_0}([0,1]^{d_0})$).
(1) Let the NPIV model (2.1) be the maintained hypothesis. We can test inequality restrictions defined by (IR) and a semiparametric structure simultaneously. For example, we can test for a partial linear structure with an increasing function g by setting $\mathcal{H}^{\text{R}} = \{h \in \Theta \oplus \mathcal{G} : \partial_{x_2} g \geq 0\}$. We can also test for the nonnegativity of the coefficient θ and a partial linear restriction by setting $\mathcal{H}^{\text{R}} = \{h \in \Theta \oplus \mathcal{G} : \partial_{x_1} h \geq 0\}$. We allow for semiparametric restriction of (ER) by taking $\mathcal{H}^{\text{R}} = \Theta \oplus \mathcal{G}$.
(2) Let the partial linear IV model be the maintained hypothesis in model (2.1) with $\mathcal{H} = \Theta \oplus \mathcal{G}$. Monotonicity in all arguments of h can be imposed by $\mathcal{H}^{\text{R}} = \{h \in \Theta \oplus \mathcal{G} : \theta \geq 0, \partial_{x_2} g \geq 0\}$. We also allow for second or higher order derivatives in the hypotheses considered above. When we test simultaneously for shape restrictions and a semiparametric structure, the reduced dimensionality of the (null) restricted model implies that Assumption 5(ii) holds, as long as $p_0 \geq p$, see Remark 4.2. In particular we may assume $d_x = 1$ if a partial linear structure is imposed as the maintained hypothesis.

4.3. Confidence Sets in L^2

One can construct L^2- confidence sets by inverting our data-driven tests for a NPIV function. The resulting confidence sets impose conjectured restrictions on the function of interest h. The $(1-\alpha)-$confidence set for a NPIV function h is given by

$$\mathcal{C}_n(\alpha) = \left\{ h \in \mathcal{H}^{\text{R}} : \frac{n\widehat{D}_J(h)}{\widehat{v}_J} \leq \widehat{\eta}_J(\alpha) \quad \text{for all } J \in \widehat{\mathcal{I}}_n \right\}. \tag{4.9}$$

This confidence set is fully data-driven and does not depend on additional tuning parameters. The following corollary exploits our previous results to characterize the asymptotic size and power properties of our procedure.

Corollary 4.2. *Let Assumptions 1(i)-(iii), 2(i)(v), 4, and 5 be satisfied. Let r_n be the rate of testing given in Theorem 4.2. Then, for any $\alpha \in (0,1)$ it holds*

$$\limsup_{n\to\infty} \sup_{h\in\mathcal{H}_0} \mathrm{P}_h \left(h \notin \mathcal{C}_n(\alpha) \right) \leq \alpha \tag{4.10}$$

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

$$\liminf_{n\to\infty} \inf_{h\in\mathcal{H}_1(\delta^\circ, r_n)} \mathrm{P}_h \left(h \notin \mathcal{C}_n(\alpha) \right) \geq 1 - \alpha. \tag{4.11}$$

Corollary 4.2 result (4.10) shows that the L^2 confidence set $\mathcal{C}_n(\alpha)$ controls size uniformly over the class of functions $\mathcal{H}_0$. Moreover, result (4.11) establishes power uniformly over the class $\mathcal{H}_1(\delta^\circ, r_n)$. We immediately see from Corollary 4.2 that the diameter of the L^2- confidence ball, $\mathsf{diam}(\mathcal{C}_n(\alpha)) = \sup \left\{ \|h_1 - h_2\|_\mu : h_1, h_2 \in \mathcal{C}_n(\alpha) \right\}$, depends on the degree of ill-posedness captured by the singular value s_{J_0}.

Corollary 4.3. *Let Assumptions 1(i)-(iii), 2(i)(v), 4, and 5 be satisfied. Then, for any $\alpha \in (0,1)$ we have*

$$\limsup_{n\to\infty} \sup_{h\in\mathcal{H}_0} \mathrm{P}_h \left(\mathsf{diam}(\mathcal{C}_n(\alpha)) \geq C J_0^{-p/d_x} \right) = 0,$$

for some constant $C > 0$ and where the optimal sieve dimension parameter J_0 is given in (4.3).

Corollary 4.3 yields a confidence set whose diameter shrinks to zero at the adaptive optimal-testing rate (of the order J_0^{-p/d_x}) and whose implementation does not require specifying the values of any unknown regularity parameters. Our confidence set $\mathcal{C}_n(\alpha)$ thus adapts to the unknown smoothness p of the unrestricted NPIV functions.[5]

Let $\mathcal{H}_0 = \{h \in B_{2,2}^{p_0} : \|h\|_{B_{2,2}^{p_0}} \leq L_0\}$ with $p_0 > p$ and $L_0 \leq L$. It is known in statistical Gaussian White noise and regression models (see Robins and Van Der Vaart [2006] and Cai and Low [2006]) that rate adaption is only possible over submodels $\mathcal{H}_0$ such that the rate of estimation over the submodel is strictly larger than the rate of testing over the "supermodel" $\mathcal{H}$. This suggests that it is impossible to adapt to the smoothness p_0 for severely ill-posed NPIV models. In the mildly ill-posed case, this leads to the restriction $n^{-p/(2(p+a)+d_x/2)} < n^{-p_0/(2(p_0+a)+d_x)}$. This condition translates into the smoothness restriction $p_0 < p\, c_a$ where $c_a = (4a + 2d_x)/(4a + d_x)$ and hence, requires $p_0 \in (p, c_a p)$. The constant c_a is close to one even under modest degrees of ill-posedness. Consequently, the gain from adaptation with respect to the smoothness of restricted classes of NPIV functions is very limited.

---

[5] We thank an anonymous referee for pointing this out.

5. Monte Carlo Studies

This section presents Monte Carlo performance of our adaptive test for monotonicity and parametric form of a NPIV function using simulation designs based on Chernozhukov et al. [2015]. See the online Appendix C for additional simulation results using other designs. In all the simulation studies we test hypotheses on the NPIV function h over the whole support of X by setting the weight function $\mu = 1$ in the implementation of our adaptive test statistic. All the simulation results are based on 5000 Monte Carlo replications for every experiment.

For all the Monte Carlo designs in this section, Y is generated according to the NPIV model (2.1) for scalar-valued random variables X and W. We let $X_i = \Phi(X_i^*)$ and $W_i = \Phi(W_i^*)$ where Φ denotes the standard normal distribution function, and generate the random vector (X_i^*, W_i^*, U_i) according to

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

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). While Chernozhukov et al. [2015] fixed $\xi = 0.5$ in their design, we let $\xi \in \{0.3, 0.5, 0.7\}$ in our simulation studies. The functional form of h varies in different Monte Carlo designs below.

5.1. Adaptive Testing for Monotonicity

We generate Y according to (2.1) and (5.1), using h from Chernozhukov et al. [2015] design:

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

This function $h(x)$ is decreasing in x, where c_0 captures the degree of monotonicity. We note that $h(x) \approx 0$ for c_0 close to zero and $h(x) \approx \phi(0)(1 - 2x)$ for c_0 close to one, where ϕ denotes the standard normal probability density function. The null hypothesis is that the NPIV function h is weakly decreasing on the support of X.

We implement our adaptive test statistic $\widehat{T}_n$ given in (2.9) 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 instrument sieve $b^{K(J)}(W)$ we also use quadratic B-spline functions with a larger number of knots with $K(J) = 2J$ or $K(J) = 4J$. Implementation of the restricted sieve NPIV estimator $\widehat{h}_J^{\text{R}}$ is straightforward using the R package `coneproj`.

n	c_0	ξ	$\widehat{\mathtt{T}}_n$ with $K=2J$			$\widehat{J}$	$\widehat{\mathtt{T}}_n$ with $K=4J$			$\widehat{J}$	$\mathtt{T}_{n,3}^{B}$ with $K=2J=6$		
			10%	5%	1%	at 5%	10%	5%	1%	at 5%	10%	5%	1%
500	0.01	0.3	0.029	0.008	0.000	3.00	0.053	0.021	0.002	3.02	0.026	0.004	0.001
		0.5	0.043	0.014	0.000	3.31	0.049	0.019	0.002	3.35	0.061	0.021	0.002
		0.7	0.047	0.021	0.003	3.56	0.049	0.024	0.006	3.57	0.048	0.016	0.002
	0.1	0.3	0.024	0.005	0.000	3.00	0.045	0.015	0.001	3.03	0.021	0.004	0.000
		0.5	0.033	0.007	0.000	3.34	0.036	0.012	0.001	3.38	0.050	0.016	0.001
		0.7	0.033	0.014	0.001	3.65	0.035	0.016	0.003	3.63	0.034	0.010	0.001
	1	0.3	0.017	0.004	0.000	3.00	0.031	0.008	0.000	3.03	0.015	0.002	0.000
		0.5	0.019	0.004	0.000	3.38	0.020	0.006	0.000	3.41	0.031	0.011	0.000
		0.7	0.015	0.005	0.000	3.76	0.017	0.007	0.001	3.74	0.015	0.004	0.000
1000	0.01	0.3	0.034	0.009	0.000	3.01	0.051	0.018	0.001	3.06	0.041	0.012	0.001
		0.5	0.034	0.013	0.001	3.49	0.043	0.016	0.002	3.44	0.067	0.026	0.003
		0.7	0.049	0.021	0.003	3.84	0.052	0.025	0.003	3.94	0.049	0.014	0.001
	0.1	0.3	0.029	0.007	0.000	3.01	0.042	0.014	0.001	3.07	0.036	0.010	0.001
		0.5	0.024	0.009	0.000	3.55	0.031	0.011	0.001	3.48	0.051	0.018	0.003
		0.7	0.031	0.014	0.002	3.99	0.037	0.016	0.003	4.08	0.027	0.008	0.001
	1	0.3	0.020	0.004	0.000	3.02	0.027	0.006	0.000	3.07	0.024	0.006	0.000
		0.5	0.012	0.002	0.000	3.63	0.015	0.004	0.000	3.54	0.029	0.008	0.001
		0.7	0.013	0.005	0.001	4.19	0.015	0.006	0.001	4.28	0.007	0.002	0.000
5000	0.01	0.3	0.035	0.012	0.001	3.38	0.041	0.015	0.001	3.38	0.076	0.028	0.003
		0.5	0.056	0.023	0.003	3.53	0.058	0.024	0.005	3.62	0.070	0.029	0.005
		0.7	0.053	0.029	0.006	4.09	0.058	0.032	0.005	4.16	0.042	0.013	0.002
	0.1	0.3	0.028	0.008	0.001	3.40	0.033	0.012	0.001	3.39	0.060	0.021	0.003
		0.5	0.035	0.012	0.001	3.67	0.036	0.014	0.002	3.75	0.040	0.013	0.002
		0.7	0.040	0.018	0.005	4.41	0.036	0.018	0.004	4.44	0.012	0.003	0.000
	1	0.3	0.015	0.004	0.000	3.48	0.017	0.006	0.000	3.40	0.035	0.012	0.001
		0.5	0.012	0.003	0.000	3.88	0.012	0.005	0.001	3.93	0.010	0.003	0.000
		0.7	0.012	0.006	0.001	4.77	0.010	0.004	0.001	4.77	0.001	0.000	0.000

Table 1: Testing Monotonicity – Empirical size of our adaptive test $\widehat{\mathtt{T}}_n$ and of the nonadaptive bootstrap test $\mathtt{T}_{n,3}^{B}$. Design from Section 5.1.

Table 1 presents the average data-driven choice of tuning parameter J, denoted by $\widehat{J}$, at the nominal level $\alpha = 0.05$. Specifically, $\widehat{J}$ is the average choice of J that maximizes $\widehat{\mathcal{W}}_J(\alpha)$ over the RES index set $\widehat{\mathcal{I}}_n$ when the null is not rejected; and is the smallest $J \in \widehat{\mathcal{I}}_n$ such that $\widehat{\mathcal{W}}_J(\alpha) > 1$ when the null is rejected. This data-driven choice of J corresponds to *early stopping* when the null is rejected. 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 NPIV 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 to reject the null hypothesis. We also see little difference between the choices $K(J) = 2J$ and $K(J) = 4J$, especially so for larger sample sizes. This is consistent with our theory that J, the dimension of the sieve basis used to approximate the unrestricted NPIV function h, is the key tuning parameter in our minimax rate adaptive testing.

Table 1 reports empirical rejection probabilities under the null hypothesis using our

adaptive test $\widehat{\mathsf{T}}_n$ for different nominal levels. 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 our adaptive test $\widehat{\mathsf{T}}_n$ provides adequate size control across different instrument strength ξ and degree of monotonicity c_0. Table G in the online Appendix C shows that our test controls size also for a sample size of $n = 10000$. In Table 1, we also compare our adaptive test to a nonadaptive bootstrap test $\mathsf{T}_{n,3}^{\mathsf{B}}$ of Fang and Seo [2021] with fixed choices of sieve dimension $J = 3$ and $K = 2J = 6$. Their statistic $\mathsf{T}_{n,3}^{\mathsf{B}}$ is computed using a standard Gaussian multiplier bootstrap critical values $\widehat{\eta}_J(\alpha)$ (and their other recommended tuning parameters of $c_n = (\log J)^{-1}$ and $\gamma_n = 0.01/\log n$) with $J = 3$. In our simulations, we use 200 bootstrap iterations. Both our adaptive tests and their nonadaptive bootstrap test for null of monotonicity are similarly undersized.

We next examine the rejection probabilities of our adaptive test when the data is generated according to (2.1) and (5.1) using the NPIV function

$$h(x) = -x/5 + c_A \left(x^2 + c_B \sin(2\pi x) \right), \tag{5.2}$$

where $0 \leq c_A \leq 2$ and $c_B \in \{0, 0.5, 1\}$. When $c_B = 0$, the null hypothesis of weakly decreasing NPIV function h over the support of X is satisfied only if $c_A \leq 0.1$. When $c_B = 0.5$ (or $c_B = 1$), the null hypothesis is satisfied only if $c_A \leq 0.1/(1 + \pi/2) \approx 0.04$ (or $c_A \leq 0.1/(1 + \pi) \approx 0.02$).

Figure 1 depicts the size-adjusted empirical power function of our adaptive monotonicity test $\widehat{\mathsf{T}}_n$ (solid lines) with $K(J) = 4J$ under the 5% nominal level for different parameters $\xi \in \{0.5, 0.7\}$ and sample sizes $n \in \{500, 1000\}$.[6] It shows that our adaptive test becomes more powerful, for $c_A > 0.1$, as the parameter of instrument strength ξ and the sample size n increase. Figure 1 also plots the size-adjusted empirical power of the nonadaptive bootstrap test $\mathsf{T}_{n,3}^{\mathsf{B}}$ (dashed lines), under the 5% nominal level, with fixed sieve dimension $J = 3$ and $K = 4J = 12$, based on 200 bootstrap iterations.

Figure 1 highlights the importance of adaptation for the power of nonparametric tests. When the alternative is of a simple quadratic form then there is little difference between our adaptive test $\widehat{\mathsf{T}}_n$ and the nonadaptive bootstrap test $\mathsf{T}_{n,3}^{\mathsf{B}}$ (with fixed sieve dimension $J = 3$). But as the amount of nonlinearity increases with the constant c_B, the nonadaptive bootstrap test becomes much less powerful than our adaptive test. For a fixed dimension parameter J, a test can have high power in a certain direction but might not be capable of detecting other nonlinearities. Note that when $c_B = 1$ the power of the nonadaptive bootstrap test $\mathsf{T}_{n,3}^{\mathsf{B}}$ does not even exceed the 5% nominal level for any value of $0 \leq c_A \leq 2$ (even for sample

---

[6]We note that size adjustment only has a very minor effect on the power curves and hence, power curves without size adjustment are not reported here due to the lack of space. The finite-sample power of our test with $K(J) = 2J$ is slightly smaller than that with $K = 4J$ when $n = 500$, but the power difference disappears when $n = 1000$.
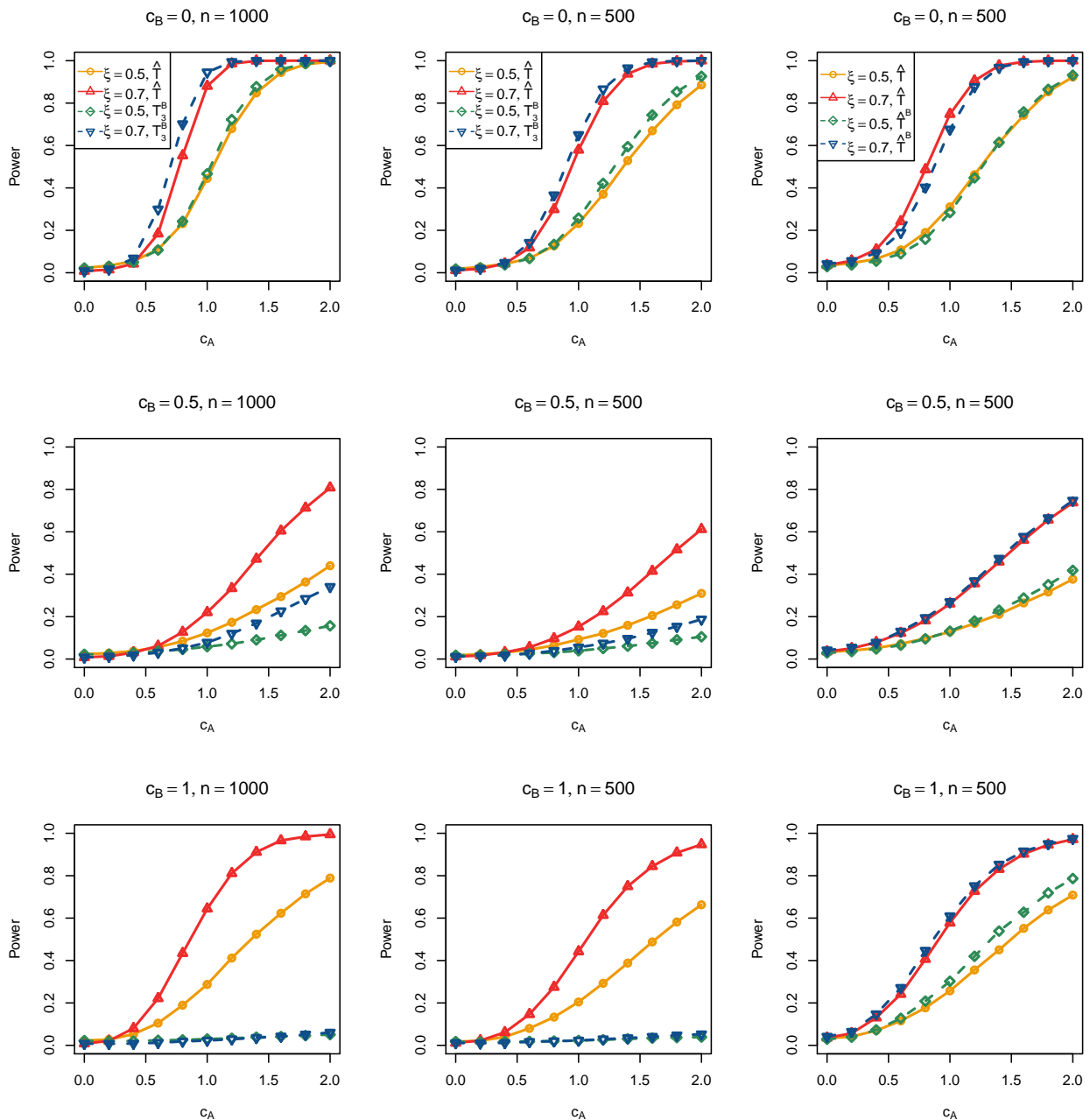
Figure 1: Testing Monotonicity – Size-adjusted empirical powers of our adaptive test $\widehat{\mathsf{T}}_n$ (solid lines), of nonadaptive bootstrap test $\mathsf{T}^{\mathsf{B}}_{n,3}$ (dashed lines in left panel with $n = 1000$ and middle panel with $n = 500$), and of "adaptive" bootstrap test $\widehat{\mathsf{T}}^{\mathsf{B}}_n$ (dashed lines in right panel with $n = 500$), $\xi = \{0.5, 0.7\}$. Design from Section 5.1.

size $n = 1000$). Similar size and power patterns are also present using another simulation design based on Chetverikov and Wilhelm [2017]; see the online Appendix C for details.

We implement the "adaptive" bootstrap analog $\widehat{\mathsf{T}}^{\mathsf{B}}_n$ of our adaptive test with $K(J) = 4J$, which requires the computation of the bootstrap test value for each J in the RES index set $\widehat{\mathcal{I}}_n$. The statistic $\widehat{\mathsf{T}}^{\mathsf{B}}_n$ also builds on bootstrap critical values $\widehat{\eta}_J(\alpha)$ following Fang and Seo [2021] (with their tuning parameters of $c_n = (\log J)^{-1}$ and $\gamma_n = 0.01/\log n$), but is now

computed for each $J \in \widehat{\mathcal{I}}_n$. As above, we use 200 bootstrap iterations. The right panel of Figure 1 shows the empirical power comparison of the "adaptive" bootstrap test $\widehat{\mathsf{T}}_n^{\mathsf{B}}$ versus our simple adaptive test $\widehat{\mathsf{T}}_n$. This indicates that the choice of sieve dimension J is the key tuning parameter for powerful inference in NPIV models.[7]

5.2. Testing for Parametric Restrictions

We now test for a parametric specification. We assume that the data is generated according to the design (2.1) and (5.1) with the NPIV function h given by (5.2). The null hypothesis is h being linear (i.e., $c_A = c_B = 0$).

We implement our adaptive test $\widehat{\mathsf{T}}_n$ given in (2.9) using quadratic B-spline basis functions with varying number of knots and where the constrained function coincides with the parametric 2SLS estimator. The number of knots varies within the RES index set $\widehat{\mathcal{I}}_n$ as implemented in the last subsection, with $K(J) = 2J$ and $K(J) = 4J$. We compare our adaptive test to the asymptotic t-test and the test by Horowitz [2006] (denoted by JH).[8] To compute the JH test that involves kernel density estimation, we follow Horowitz [2006] to estimate the joint density f_{XW} using the kernel $K(v) = (15/16)(1 - v^2)^2 \mathbb{1}\{|v| \leq 1\}$ with the kernel bandwidth chosen via cross-validation that minimizes mean squared error of density estimation.

n	ξ	$\widehat{\mathsf{T}}_n$ with $K = 2J$	$\widehat{J}$	$\widehat{\mathsf{T}}_n$ with $K = 4J$	$\widehat{J}$	t-test	JH test
500	0.3	0.010	3.00	0.023	3.03	0.001	0.053
	0.5	0.023	3.34	0.030	3.50	0.024	0.057
	0.7	0.030	3.61	0.032	3.63	0.042	0.054
1000	0.3	0.013	3.01	0.023	3.07	0.005	0.055
	0.5	0.020	3.52	0.030	3.50	0.038	0.055
	0.7	0.036	3.91	0.039	4.00	0.049	0.056
5000	0.3	0.022	3.38	0.028	3.41	0.029	0.057
	0.5	0.039	3.59	0.042	3.64	0.048	0.056
	0.7	0.045	4.18	0.048	4.18	0.050	0.056

Table 2: Testing Parametric Form – Empirical size of our adaptive test $\widehat{\mathsf{T}}_n$, the t-test and JH test. 5% nominal level. Design from Section 5.2.

Table 2 reports empirical rejection probabilities under the null hypothesis of linearity of h, of the tests at the 5% nominal level. Results are presented under different strength of instrument values for $\xi \in \{0.3, 0.5, 0.7\}$. Overall, our adaptive test $\widehat{\mathsf{T}}_n$ provides adequate size control for different parameter values of ξ. Table H in the online Appendix C shows

---

[7]In the first submitted version Breunig and Chen [2020], we present evidences from simulations and real data applications that, in terms of finite-sample size and power, the performances of our easy-to-compute adaptive test and its bootstrapped version are virtually the same. We no longer report those results here due to the lack of space.

[8]Horowitz [2006] already demonstrated in his simulation studies that his test is more powerful than several existing tests including Bierens [1990]'s.

that our test controls size also for a sample size of $n = 10000$. As the sample size grows to $n = 5000$, the difference in empirical size of our adaptive tests between $K(J) = 2J$ and $K(J) = 4J$ is only minor. This is consistent with our theory that the sieve dimension J used to approximate nonparametric alternative NPIV function is the key tuning parameter in our adaptive, minimax rate-optimal test. Table 2 also reveals that the JH test is slightly over-sized while our test and the t-test are under-sized in small samples ($n = 500, 1000$).
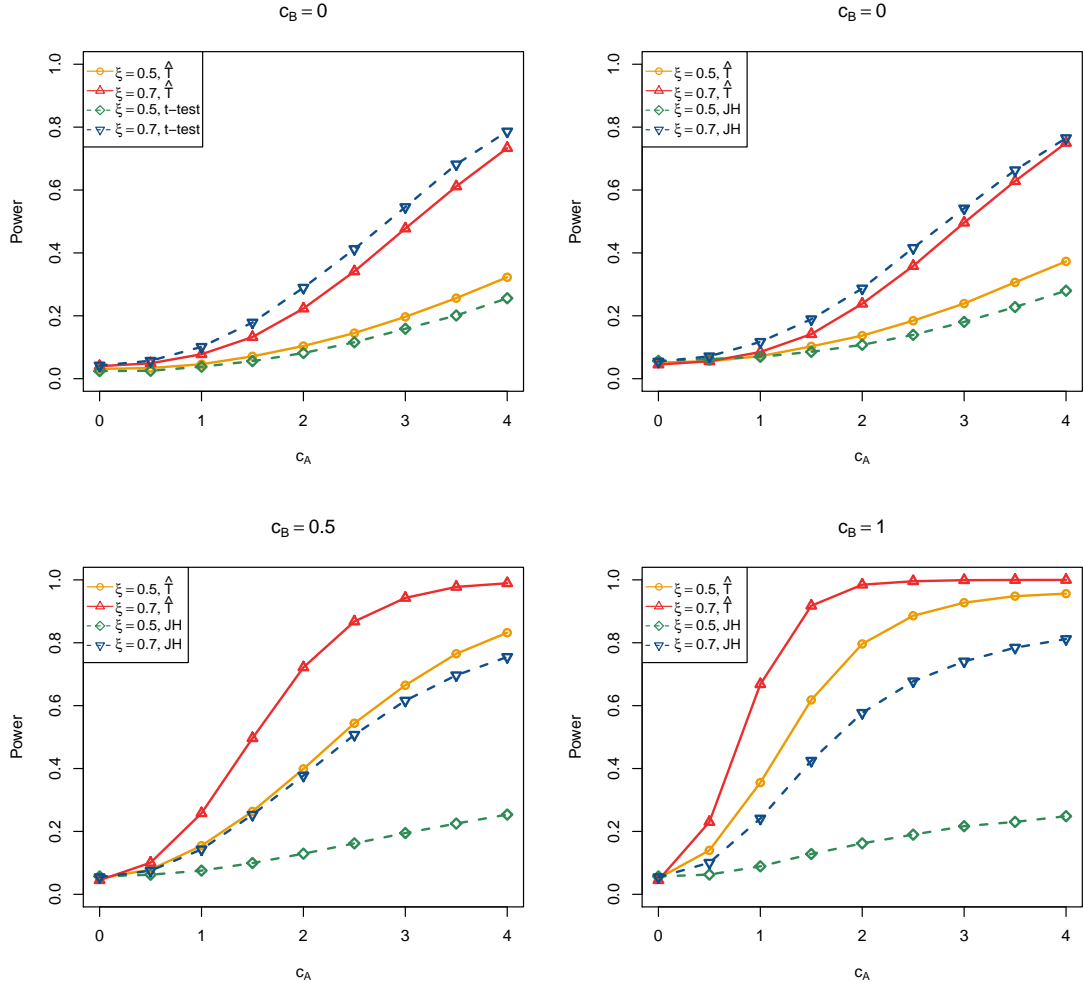
Figure 2: Testing Parametric Form – Size-adjusted empirical power of our adaptive test $\widehat{T}_n$ (solid lines), $\xi = \{0.5, 0.7\}$, $n = 500$. First row: Power comparison to the t-test and JH test (dashed lines) when $c_B = 0$; Second row: Power comparison to the JH test when $c_B = 0.5$ and $c_B = 1$. Design from Section 5.2.

Figure 2 provides size-adjusted empirical power curves for the 5% level tests with sample size $n = 500$. See Figure C in the online Appendix C for empirical power curves with a larger sample size $n = 1000$. From both figures, we see that our adaptive test $\widehat{T}_n$ (solid lines) with $K(J) = 4J$ has power similar to the asymptotic t-test and the JH test (dashed lines) for a simple alternative with $c_B = 0$ (i.e., the alternative function is quadratic). When the alternative function becomes more complex with $c_B = 0.5$, our adaptive test

becomes more powerful than the JH test. The difference in power is enlarged as the degree of nonlinearity increases to $c_B = 1$. This is theoretically sensible since Horowitz [2006] test is designed to have power against $n^{-1/2}$ smooth alternative only. To sum up, our adaptive minimax test not only controls size, but also has very good finite-sample power uniformly against a large class of nonparametric alternatives.

Finally in online Appendix C we present additional simulation comparisons of our adaptive test against our adaptive version of Bierens [1990]'s type test when the dimension of conditional instrument W is larger than the dimension of the endogenous variables X. We observe that our adaptive test again have very good size control and even better finite-sample power when $d_w > d_x$.

6. Empirical Applications

We present two empirical applications of our adaptive test for NPIV models. The first one tests for connected substitutes restrictions in differentiated products demand using market level data. The second one tests for monotonicity, convexity or parametric specification of Engel curves for non-durable good consumption using household level data.

In both empirical applications, we implement our adaptive test $\widehat{\mathsf{T}}_n$ given in (2.9) with $\mu \equiv 1$ and $K(J) = 4J$. The null hypothesis is rejected at the nominal level $\alpha = 0.05$ whenever $\widehat{\mathcal{W}}_J(\alpha) > 1$ for some $J \in \widehat{\mathcal{I}}_n$ (the RES index set). Tables in this section report a set $\widehat{\mathcal{J}} \subset \widehat{\mathcal{I}}_n$, which equals to $\arg\max_{J \in \widehat{\mathcal{I}}_n} \widehat{\mathcal{W}}_J(\alpha)$ when our test fails to reject the null hypothesis and equals to $\{J \in \widehat{\mathcal{I}}_n : \widehat{\mathcal{W}}_J(\alpha) > 1\}$ when our test rejects the null. Below, we report $\widehat{\mathcal{W}}_{\widehat{J}}$ with $\widehat{J}$ being the minimal integer of $\widehat{\mathcal{J}}$. We also report the corresponding p value, which should, by Bonferroni correction, be compared to the nominal level $\alpha = 0.05$ divided by the cardinality of $\widehat{\mathcal{I}}_n$. Finally, since our test is based on a leave-one-out version, the value of $\widehat{\mathcal{W}}_{\widehat{J}}$ could be negative.

6.1. Adaptive Testing for Connected Substitutes in Demand for Differential Products

Recently Berry and Haile [2014] provide conditions under which a nonparametric demand system for differentiated products can be inverted to NPIV equations using Market level data. A key restriction is what they call "connected substitutes". Compiani [2021] applies their nonparametric identification results and estimates the system of inverse demand by directly imposing the connected substitutes restrictions in his implementation of sieve NPIV estimator, and obtains informative results as an alternative to BLP demand in simulation studies and a real data application.

We revisit Compiani [2021]'s empirical application using the 2014 Nielsen scanner data

set that contains market (store/week) level data of consumers in California choosing from organic strawberries, non-organic strawberries and an outside option. While Compiani [2021] directly imposes "connected substitutes" restriction in his sieve NPIV estimation of inverse demand, we want to test this restriction. Following Compiani [2021] we consider

$$X_o + U = h(\mathbf{P}, S_o, S_{no}, In), \qquad \mathrm{E}[U|\mathbf{W}_p, X_o, X_{no}, In] = 0,$$

where h denotes the inverse of the demand for organic strawberries, X_o denotes a measure of taste for organic products, X_{no} denotes the availability of other fruit, S_o and S_{no} denote the endogenous shares of the organic and non-organic strawberries, respectively. (X_o, X_{no}) are the two included instruments for the two endogenous shares (S_o, S_{no}). In denotes store-level (zip code) income and U unobserved shocks for organic produce. The vector $\mathbf{P} = (P_o, P_{no}, P_{out})$ denotes the endogenous prices of organic strawberries, non-organic strawberries, and non-strawberry fresh fruit, respectively. We follow Compiani [2021] and let $\mathbf{W}_p = (W_o, W_{no}, W_{out}, W_{s1}, W_{s2})$ be a $5-$ dimensional vector of conditional instruments for the price vector $\mathbf{P}$, including 3 Hausman-type instrumental variables (W_o, W_{no}, W_{out}) and 2 shipping-point spot prices (W_{s1}, W_{s2}) (as proxies for the wholesale prices faced by retailers).

As shown by Compiani [2021, Lemma 1], the connected substitutes assumption of Berry and Haile [2014] implies the following shape restrictions on the function h: First, h is weakly increasing in the organic product price P_o. Second, h is weakly increasing in the organic product share S_o. Third, h is weakly increasing in the non-organic product share S_{no}. Fourth, $\partial h/\partial s_o \geq \partial h/\partial s_{no}$ (the so-called diagonal dominance). Below, we test for these inequality restrictions.

We consider a subset of the data set of Compiani [2021][9], where income ranges from the first and to the third quartile of its distribution and prices for organic produces are restricted to be above its 1st and below its 99th percentile. The resulting sample has size $n = 11910$. We implement our adaptive test $\widehat{\mathsf{T}}_n$ by making use of a semiparametric specification of the function h: we consider the tensor product of quadratic B-splines $\psi^{J_1}(P_o)$ and the vector $(1, In, P_{no}, \psi^3(S_o))$, where we use a cubic B-spline transformation of S_o without knots and without intercept, hence $J = 6J_1$. The variables $(P_{out}, S_{no}, S_{no}P_{no}, S_{no}S_o)$ are included additively and we set $K(J) = 4J$. We obtain the index set $\widehat{\mathcal{I}}_n = \{24, 30, 36\}$.

According to Table 3, our adaptive test fails to reject that h is weakly increasing in the own price at the nominal level $\alpha = 0.05$, but rejects $\partial h/\partial p_o \leq 0$. Similarly, this table shows that our adaptive test also fails to reject that h is weakly increase in non-organic shares and rejects that h is weakly decreasing in S_o. When testing partial derivatives, our test fails to reject that the partial effect with respect to the non-organic share is constant.

---

[9]For details on the construction of the data and descriptive statistics, see Compiani [2021, Appendix F].

H_0	$\widehat{\mathcal{W}}_{\widehat{\mathcal{J}}}$	p val.	reject H_0?	$\widehat{\mathcal{J}}$
$\partial h/\partial p_o \geq 0$	0.714	0.036	no	$\{36\}$
$\partial h/\partial p_o \leq 0$	2.635	0.000	yes	$\{36\}$
$\partial h/\partial s_o \geq 0$	0.554	0.057	no	$\{36\}$
$\partial h/\partial s_o \leq 0$	1.786	0.002	yes	$\{24, 30, 36\}$
$\partial h/\partial s_{no} \geq 0$	-0.105	0.479	no	$\{24\}$
$\partial h/\partial s_{no} \leq 0$	-0.206	0.878	no	$\{24\}$
$\partial h/\partial s_o \geq \partial h/\partial s_{no}$	0.554	0.057	no	$\{36\}$
$\partial h/\partial s_o \leq \partial h/\partial s_{no}$	1.786	0.002	yes	$\{24, 30, 36\}$

Table 3: Adaptive testing for the shape of h.

Finally, the last two rows show that our test provides empirical evidence for the diagonal dominance restriction.

6.2. Adaptive Testing for Engel Curves

The system of Engel curves plays a central role in the analysis of consumer demand for non-durable goods. It describes the $i-$th household's budget share $Y_{\ell,i}$ for non-durable goods ℓ as a function of its log-total expenditure X_i and other exogenous characteristics such as family size and age of the head of the $i-$th household. 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] propose a popular extension of this system of linear Engel curves to include a squared term in log-total expenditure, and their parametric Student t test rejects linear form in favor of quadratic Engel curves.

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 a conditional 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": $\mathrm{E}[Y_\ell - h_\ell(X)|W] = 0$. We use the same quadratic B-spline basis with up to 3 knots to approximate all the Engel curves and set $K(J) = 4J$. Hence the index set $\widehat{\mathcal{I}}_n = \{3, 4, 5\}$ is the same for the different Engel curves.

Table 4 reports our adaptive test for weak monotonicity of Engel curves. It shows 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 are 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 uniform confidence bands are constructed using a sieve score bootstrapped critical values with non-data-driven choice of sieve dimension J, and 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 95% uniform confidence bands. Table 5 reports our adaptive test for convexity and concavity of these Engel curves. At the 5% nominal level, we reject convexity of travel goods and reject concavity of Engel curves for fuel consumption. These are in line with Chen and Christensen [2018, Figure 4], but again, significant statements about the convexity/concavity of Engel curves are only possible using our adaptive testing procedure. Finally, Table 6 presents our adaptive tests for linear or quadratic specifications (against nonparametric alternatives) of the Engel curves for the four goods. At the nominal level $\alpha = 0.05$, this table shows that our adaptive test fails to reject a quadratic form for all the goods, while it rejects a linear Engel curve for fuel and travel goods. Our results are consistent with the conclusions obtained by Banks et al. [1997] using Student t-test for linear against quadratic forms of Engel curves.

| | H_0: h is increasing | | | | H_0: h is decreasing | | | |
Goods	$\widehat{\mathcal{W}}_{\widehat{\mathcal{J}}}$	p value	reject H_0?	$\widehat{\mathcal{J}}$	$\widehat{\mathcal{W}}_{\widehat{\mathcal{J}}}$	p value	reject H_0?	$\widehat{\mathcal{J}}$
"food in"	2.741	0.000	yes	$\{3\}$	-0.286	1.000	no	$\{4\}$
"fuel"	7.820	0.000	yes	$\{3,4\}$	0.483	0.0375	no	$\{3\}$
"travel"	2.413	0.000	yes	$\{3\}$	0.336	0.075	no	$\{3\}$
"leisure"	0.256	0.115	no	$\{4\}$	4.346	0.000	yes	$\{3,4\}$

Table 4: Adaptive testing for monotonicity of Engel curves.

| | H_0: h is convex | | | | H_0: h is concave | | | |
Goods	$\widehat{\mathcal{W}}_{\widehat{\mathcal{J}}}$	p value	reject H_0?	$\widehat{\mathcal{J}}$	$\widehat{\mathcal{W}}_{\widehat{\mathcal{J}}}$	p value	reject H_0?	$\widehat{\mathcal{J}}$
"food in"	-0.254	1.000	no	$\{3\}$	-0.286	1.000	no	$\{3\}$
"fuel"	-0.280	1.000	no	$\{4\}$	1.431	0.000	yes	$\{3\}$
"travel"	1.049	0.003	yes	$\{3\}$	-0.275	1.000	no	$\{5\}$
"leisure"	-0.169	0.818	no	$\{5\}$	0.639	0.023	no	$\{4\}$

Table 5: Adaptive testing for convexity/concavity of Engel curves.

7. Conclusion

In this paper, we propose a new adaptive, minimax rate-optimal test on a structural function in NPIV models. We can test for equality (e.g., parametric, semiparametric forms)

Goods	H_0: h is linear				H_0: h is quadratic			
	$\widehat{\mathcal{W}}_{\widehat{\mathcal{J}}}$	p value	reject H_0?	$\widehat{\mathcal{J}}$	$\widehat{\mathcal{W}}_{\widehat{\mathcal{J}}}$	p value	reject H_0?	$\widehat{\mathcal{J}}$
"food in"	-0.273	0.922	no	{3}	0.125	0.233	no	{3}
"fuel"	1.623	0.000	yes	{3}	-0.120	0.612	no	{5}
"travel"	1.210	0.001	yes	{3}	-0.014	0.415	no	{4}
"leisure"	0.691	0.074	no	{4}	0.513	0.041	no	{4}

Table 6: Adaptive testing for linear/quadratic specification of Engel curves.

and polyhedral cone type inequality (e.g., monotonicity, convexity, concavity, supermodularity and other shapes) restrictions. Our test statistic is based on a leave-one-out sieve estimator of the quadratic distance between restricted and unrestricted structural functions in NPIV models. We establish the minimax rate of testing against classes of nonparametric alternative models. The minimax rate of testing depends on the optimal choice of tuning parameters that in turn depend on unknown model features. We then provide computationally simple data-driven choices of sieve tuning parameters. The resulting test attains the optimal minimax rate of testing, adapts to the unknown smoothness of nonparametric alternative NPIV functions, and is robust to the unknown degree of endogeneity and strength of instruments. Data-driven confidence sets (in L^2) can be obtained by inverting our adaptive test. Monte Carlo studies demonstrate that our simple, adaptive test has good size and power properties in finite samples for testing monotonicity and parametric restrictions in NPIV models, without the need of using computationally intensive bootstrapped critical values. Empirical applications illustrate the power of our adaptive tests for checking shape restrictions (connected substitutes, monotonicity, convexity) and parametric specifications in nonparametric models with endogeneity.

References

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. Specification testing in nonparametric instrumental quantile regression. *Econometric Theory*, 36(4):583–625, 2020.

C. Breunig and X. Chen. Adaptive, rate-optimal testing in instrumental variables models. *arXiv preprint arXiv:2006.09587v1*, 2020.

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

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.

X. Chen, T. Christensen, and S. Kankanala. Adaptive estimation and uniform confidence bands for nonparametric iv. *arXiv preprint arXiv:2107.11869*, 2021.

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

D. Chetverikov. Testing regression monotonicity in econometric models. *Econometric Theory*, 35(4):729–776, 2019.

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

D. Chetverikov, A. Santos, and A. M. Shaikh. The econometrics of shape restrictions. *Annual Review of Economics*, 10:31–63, 2018.

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. *Quantitative Economics*, forthcoming, 2021.

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

Z. Fang and J. Seo. A projection framework for testing shape restrictions that form convex cones. *Econometrica*, 89(5):2439–2458, 2021.

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

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.

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

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

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.

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.

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

T. Inglot. Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30(2):339–351, 2010.

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

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, pages 1302–1338, 2000.

R. L. Matzkin. Restrictions of economic theory in nonparametric methods. *Handbook of econometrics*, 4:2523–2558, 1994.

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.

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

J. Tao. Trinity tests of functions for conditional moment models. *Journal of Multivariate Analysis*, 178:104604, 2020.

H. Triebel. Theory of function spaces iii, volume 100 of. *Monographs in Mathematics*, 2006.

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

A. Proofs of Minimax Testing Results in Section 3

Proof of Theorem 3.1. Let P_θ denote the joint distribution of (Y, X, W) where $Y = Th_\theta + V$ where $V|W \sim \mathcal{N}(0, \sigma^2(W))$, the so called reduced-form NPIR as in Chen and Reiß [2011]. It is sufficient to consider the scalar case $d_x = 1$ (otherwise consider tensor product wavelet bases as in Chen and Christensen [2018, Appendix E]). Without loss of generality, we may assume that the support of μ coincides with the interval $[0, 1]$. We may choose a subset $\mathcal{M}$ of $\{0, \ldots, 2^j - 1\}$ with $\#(\mathcal{M}) \sim 2^j$ such that $\widetilde{\psi}_{j,m}$ and $\widetilde{\psi}_{j,m'}$ have disjoint support for each $m, m' \in \mathcal{M}$ with $m \neq m'$. For any function $h_0 \in \mathcal{H}_0$ we set $\vartheta = (\vartheta_m)_{m \in \mathcal{M}}$ with $\vartheta_m \in \{-1, 1\}$ and define

$$h_\vartheta = h_0 + \frac{\sqrt{\delta_*} 2^{-jp}}{\sqrt{\#(\mathcal{M})}} \sum_{m \in \mathcal{M}} \vartheta_m \widetilde{\psi}_{j,m}$$

for some $\delta_* > 0$. Therefore, for CDV wavelet basis functions $\widetilde{\psi}_{j,l}$, which are orthonormal in $L^2[0, 1]$, there exists a constant $C > 0$ such that

$$\|h_\vartheta - h_0\|_{B^p_{2,2}} \leq C \sum_{j,m} 2^{2jp} \int_0^1 (h_\vartheta - h_0)^2(x) \widetilde{\psi}^2_{j,m}(x) dx = \frac{C\delta_*}{\#(\mathcal{M})} \sum_{m \in \mathcal{M}} = C\delta_*$$

and we conclude that $h_\vartheta - h_0 \in \mathcal{H}$ for δ_* sufficiently small. Denoting $r_n = 2^{-jp}$ we derive the lower bound

$$\|h_\vartheta - h_0\|_\mu^2 \geq \int_0^1 (h_\vartheta - h_0)^2(x)dx = \frac{\delta_* 2^{-2jp}}{\#(\mathcal{M})} \sum_{m \in \mathcal{M}} = \delta_* r_n^2$$

which shows $h_\vartheta \in \mathcal{H}_1(\delta_*, r_n)$. In particular, we obtain $h_\vartheta \in \mathcal{H}_1(\delta, r_n)$ for all $0 < \delta < \delta_*$.

Let P^* denote the mixture distribution obtained by assigning weight 2^{-m} to P_ϑ for each of the 2^m realizations of ϑ. For all $0 < \delta < \delta_*$, the observation that $h_\theta \in \mathcal{H}_1(\delta, r_n)$ yields the following reduction to testing between two probability measures:

$$\inf_{T_n} \Big\{ \sup_{h \in \mathcal{H}_0} P_h(T_n = 1) + \sup_{h \in \mathcal{H}_1(\delta, r_n)} P_h(T_n = 0) \Big\}$$

$$\geq \inf_{T_n} \Big\{ P_0(T_n = 1) + \sup_{\theta \in \{\vartheta: \vartheta_m \in \{-1,1\}^m\}} P_\theta(T_n = 0) \Big\}$$

$$\geq \inf_{T_n} \big\{ P_0(T_n = 1) + P^*(T_n = 0) \big\} \geq 1 - \|P_0 - P^*\|_{TV}, \tag{A.1}$$

where the last inequality is due to the proof of Collier et al. [2017, Lemma 3] and $\|\cdot\|_{TV}$ denotes the total variation distance. Due to Assumption 1 we may apply Chen and Christensen [2018, Lemma G.8] which yields $\|P_0 - P^*\|_{TV}^2 \leq Cn^2 2^{-4jp}\nu(2^j)^4/\#(\mathcal{M})$ for some constant $C > 0$.

Consider the mildly ill-posed case ($\nu(2^j) = 2^{-ja}$). The choice $2^j = cn^{2/(4(p+a)+1)}$ for some sufficiently small $c > 0$ gives $\|P_0 - P^*\|_{TV}^2 \leq Cn^2 2^{-j(4(p+a)+1)} \leq 1 - \alpha$ and $r_n \sim 2^{-jp} \sim n^{-2p/(4(p+a)+1)}$. Consider now the severely ill-posed case ($\nu(2^j) = \exp(-\frac{1}{2}2^{-ja})$). The choice of $2^j = (c \log n)^{1/a}$ for some $c \in (0, 1)$ yields $\|P_0 - P^*\|_{TV}^2 \leq 1 - \alpha$ for n sufficiently large and $r_n \sim 2^{-jp} \sim (\log n)^{-p/a}$, which completes the proof. $\qquad\square$

Proof of Theorem 3.2. We first control the type I error of the test $T_{n,J}$ given in (3.1). Note that

$$\limsup_{n \to \infty} P_{h_0}(T_{n,J} = 1) = \limsup_{n \to \infty} P_{h_0}\left(n\widehat{D}_J(h_0) > \eta_J(\alpha)\widehat{v}_J\right) \leq \alpha$$

by Lemma E.8. To control the type II error, we calculate

$$P_h(T_{n,J} = 0) \leq P_h\left(n\widehat{D}_J(h_0) \leq \eta_J(\alpha)v_J, \quad \widehat{v}_J \leq (1 + c_0)v_J\right) + P_h\left(\widehat{v}_J > (1 + c_0)v_J\right)$$

$$= P_h\left(n\widehat{D}_J(h_0) \leq (1 + c_0)\eta_J(\alpha)v_J\right) + o(1) = o(1)$$

uniformly for $h \in \mathcal{H}_1(\delta^*, r_{n,J})$, where the second equation is due to Lemma E.6(i) and the last equation is due to Lemma B.1(i). $\qquad\square$

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

34

observe

$$s_J = \inf_{h \in \Psi_J} \frac{\|\Pi_K T h\|_{L^2(W)}}{\|h\|_\mu} \geq \inf_{h \in \Psi_J} \frac{\|T h\|_{L^2(W)}}{\|h\|_\mu} - \sup_{h \in \Psi_J} \frac{\|(\Pi_K T - T)h\|_{L^2(W)}}{\|h\|_\mu} = (1 - o(1))\tau_J^{-1}$$

by identification imposed in Assumption 1(ii) and Assumption 3, i.e., $\sup_{h \in \Psi_J} \tau_J \|(\Pi_K T - T)h\|_{L^2(W)}/\|h\|_\mu = o(1)$. Consider the mildly ill-posed case ($\tau_j \sim j^{a/d_x}$). We have $J_* \sim n^{2d_x/(4(p+a)+d_x)}$ and hence,

$$n^{-1}\tau_{J_*}^2 \sqrt{J_*} \sim n^{-1} J_*^{1/2 + 2a/d_x} \sim n^{-4p/(4(p+a)+d_x)}$$

and for the bias term $J_*^{-2p/d_x} \sim n^{-4p/(4(p+a)+d_x)}$. Consider now the severely ill-posed case ($\tau_j \sim \exp(j^{a/d_x}/2)$). The definition of J_* implies $J_* \lesssim (\log n - (4p + d_x)/(2a) \log \log n)^{d_x/a}$, which gives

$$n^{-1}\tau_{J_*}^2 \sqrt{J_*} \sim n^{-1} \sqrt{J_*} \exp(J_*^{a/d_x}) \lesssim \left(\log n - \frac{4p + d_x}{2a} \log \log n \right)^{d_x/(2a)} (\log n)^{-(4p+d_x)/(2a)}$$
$$\sim (\log n)^{d_x/(2a)} (\log n)^{-(4p+d_x)/(2a)} \sim (\log n)^{-2p/a}$$

and for the bias term $J_*^{-2p/d_x} \sim (\log n)^{-2p/a}$. □

B. Proofs of Adaptive Testing Results in Section 4

We first introduce additional notation. 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 $A = (G_b^{-1/2} S G^{-1/2})_l^- G_b^{-1/2}$ and $\widehat{A} = (\widehat{G}_b^{-1/2} \widehat{S} \widehat{G}^{-1/2})_l^- \widehat{G}_b^{-1/2}$. For all $J \geq 1$ such that $s_J = s_{\min}(G_b^{-1/2} S G^{-1/2}) > 0$ it holds

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

Let $V_i^J := (Y_i - \Pi_{\mathrm{R}} h(X_i)) A b^K(W_i)$ and $\widetilde{b}^K(\cdot) = G_b^{-1/2} b^K(\cdot)$. For any NPIV function $h \in \mathcal{H}$, we introduce its population 2SLS projection onto the sieve space Ψ_J as

$$Q_J h(\cdot) = \psi^J(\cdot)'(G_b^{-1/2}S)_l^- \mathrm{E}[h(X)\widetilde{b}^K(W)] . \tag{B.1}$$

We let $Z_i = (Y_i, X_i, W_i)$ and introduce a function

$$R(Z_i, Z_{i'}, D_i) = (Y_i - \Pi_{\mathrm{R}} h(X_i))\mathbb{1}_{D_i} b^K(W_i)' A' A b^K(W_{i'})(Y_{i'} - \Pi_{\mathrm{R}} h(X_{i'}))\mathbb{1}_{D_{i'}}$$
$$- \mathrm{E}_h[(Y - \Pi_{\mathrm{R}} h(X))\mathbb{1}_D b^K(W)]' A' A \, \mathrm{E}_h[b^K(W)(Y - \Pi_{\mathrm{R}} h(X))\mathbb{1}_D]$$

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 $M_i = \{|Y_i - \Pi_{\mathrm{R}} h(X_i)| \leq M_n\}$ and $M_n = \sqrt{n}\, \zeta_{\overline{J}}^{-1} (\log \log \overline{J})^{-3/4}$. Based on kernels R_l, where $l = 1, 2$, we introduce the U-statistic

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

We also introduce the notation

$$\Lambda_1 = \left(\frac{n(n-1)}{2} \, \mathrm{E}[R_1^2(Z_1, Z_2)] \right)^{1/2}, \qquad \Lambda_2 = n \sup_{\|\nu\|_{L^2(Z)} \leq 1, \|\kappa\|_{L^2(Z)} \leq 1} \mathrm{E}[R_1(Z_1, Z_2)\nu(Z_1)\kappa(Z_2)],$$

$$\Lambda_3 = \left(n \sup_z |\,\mathrm{E}[R_1^2(Z_1, z)]|\right)^{1/2}, \qquad \text{and} \qquad \Lambda_4 = \sup_{z_1, z_2} |R_1(z_1, z_2)|.$$

For testing equality restrictions, we let $\eta_J(\alpha) = \big(q\big(\alpha/\#(\mathcal{I}_n), J\big) - J\big)/\sqrt{J}$ denote the deterministic analog of $\widehat{\eta}_J(\alpha)$ given in (2.8). Below, we also denote by $C > 0$ a generic constant that may be different in different uses.

Proof of Theorem 4.1. We prove this result in three steps. First, we bound the type I error of the test statistic

$$\widetilde{\mathrm{T}}_n = \mathbb{1}\left\{ \max_{J \in \mathcal{I}_n} \big(n\widehat{D}_J(h_0)/(\eta_J'(\alpha)\mathrm{v}_J)\big) > 1\right\}$$

for some $\eta_J'(\alpha) > 0$. Second, we bound the type II error of $\widetilde{\mathrm{T}}_n$ where $\eta_J'(\alpha)$ is replaced by $\eta_J''(\alpha) > 0$. Let $\eta_J'(\alpha)$ and $\eta_J''(\alpha)$ be such that $\eta_J(\alpha) = \eta_J'(\alpha)/(1 - c_0) = \eta_J''(\alpha)/(1 + c_0)$ for some constant $0 < c_0 < 1$. Finally, we show that the derived bounds in the previous steps are sufficient to control the type I and II error of our test statistic $\widehat{\mathrm{T}}_n$ for simple null hypotheses.

Step 1: To control the first type error of the test statistic $\widetilde{\mathrm{T}}_n$, we make use of the decomposition under $\mathcal{H}_0 = \{h_0\}$:

$$\mathrm{P}_{h_0}\big(\widetilde{\mathrm{T}}_n = 1\big) \leq \mathrm{P}_{h_0}\Big(\max_{J \in \mathcal{I}_n} \Big| \frac{1}{\eta_J'(\alpha)\mathrm{v}_J(n-1)} \sum_{j=1}^{J} \sum_{i \neq i'} V_{ij} V_{i'j}\Big|$$

$$+ \max_{J \in \mathcal{I}_n} \Big| \frac{1}{\eta_J'(\alpha)\mathrm{v}_J(n-1)} \sum_{i \neq i'} U_i U_{i'} b^K(W_i)'\Big(A'A - \widehat{A}'\widehat{A}\Big) b^K(W_{i'})\Big| > 1\Big)$$

$$\leq \underbrace{\mathrm{P}_{h_0}\Big(\max_{J \in \mathcal{I}_n} \big| n\mathcal{U}_{J,1}/(\eta_J'(\alpha)\mathrm{v}_J)\big| > \frac{1}{4}\Big)}_{=I} + \underbrace{\mathrm{P}_{h_0}\Big(\max_{J \in \mathcal{I}_n} \big| n\mathcal{U}_{J,2}/(\eta_J'(\alpha)\mathrm{v}_J)\big| > \frac{1}{4}\Big)}_{=II}$$

$$+ \underbrace{\mathrm{P}_{h_0}\Big(\max_{J \in \mathcal{I}_n} \Big| \frac{1}{\eta_J'(\alpha)\mathrm{v}_J(n-1)} \sum_{i \neq i'} U_i U_{i'} b^K(W_i)'\big(A'A - \widehat{A}'\widehat{A}\big) b^K(W_{i'})\Big| > \frac{1}{2}\Big)}_{=III},$$

36

using the notation $U_i = Y_i - h_0(X_i)$. Consider I. From Lemma F.1 and Lemma F.2 with $M_n = \sqrt{n}\,\zeta_{\overline{J}}^{-1}(\log\log\overline{J})^{-3/4}$ we infer for all $J \in \mathcal{I}_n$ that

$$
\begin{aligned}
\Lambda(u, J) &:= \Lambda_1\sqrt{u} + \Lambda_2 u + \Lambda_3 u^{3/2} + \Lambda_4 u^2 \\
&\leq n v_J\sqrt{u/2} + \overline{\sigma}^2 n s_J^{-2} u + \overline{\sigma}^2 n s_J^{-2}(\log\log\overline{J})^{-3/4} u^{3/2} + n s_J^{-2}(\log\log\overline{J})^{-3/2} u^2
\end{aligned}
$$

for n sufficiently large. Replacing in the previous inequality u by $u_J = 2\log\log J^{c_\alpha}$ where $c_\alpha = \sqrt{1 + (\pi/\log 2)^2}/\sqrt{\alpha}$, we obtain for n sufficiently large:

$$
\begin{aligned}
\Lambda(u_J, J) &\leq n v_J\sqrt{\log\log J^{c_\alpha}} + \frac{2\overline{\sigma}^2 n}{s_J^2}\log\log J^{c_\alpha} + \frac{\overline{\sigma}^2 n}{s_J^2}(2\log\log J^{c_\alpha})^{3/4} + \frac{4n}{s_J^2}\sqrt{\log\log J^{c_\alpha}} \\
&\leq \frac{5}{4}n v_J\sqrt{\log\log J - \log\alpha} + 3\overline{\sigma}^2 n s_J^{-2}(\log\log J - \log\alpha) \\
&\leq \frac{5}{1-c_0}n v_J\eta_J'(\alpha) + \frac{12\overline{\sigma}^2}{1-c_0}n s_J^{-2}\eta_J'(\alpha)\sqrt{\log\log J},
\end{aligned}
$$

where the second inequality follows from Lemma E.9, that is, $\sqrt{\log\log J - \log\alpha} \leq 4\eta_J'(\alpha)/(1-c_0)$. Assumption 4(ii) implies for all $J \in \mathcal{I}_n$ and for n sufficiently large that

$$
\Lambda(u_J, L(J)) \leq \frac{n-1}{2}v_J\eta_J'(\alpha)
$$

where $L(J) = \exp(1/6)J\underline{J}^{-1/2}$ and using $s_J^{-2} \leq \underline{\sigma}^{-2}v_J$ by Lemma E.2. Consequently, the exponential inequality for degenerate U-statistics in Lemma F.1 with $u = 2\log\log J^{c_\alpha}$ together with the definition of $\mathcal{I}_n$, i.e., $J = \underline{J}2^j$ for all $J \in \mathcal{I}_n$, yields for n sufficiently large:

$$
\begin{aligned}
I &\leq \sum_{J \in \mathcal{I}_n} \mathrm{P}_{h_0}\left(|n\mathcal{U}_{J,1}| > \frac{\eta_J'(\alpha)}{4}v_J\right) = \sum_{J \in \mathcal{I}_n} \mathrm{P}_{h_0}\left(\left|\sum_{i<i'} R_1(Z_i, Z_{i'})\right| \geq \frac{\eta_J'(\alpha)}{4}\frac{n-1}{2}v_J\right) \\
&\leq 6\sum_{J \in \mathcal{I}_n} \exp\left(-2\log\log(L(J)^{c_\alpha})\right) = 6\,c_\alpha^{-2}\sum_{J \in \mathcal{I}_n}\left(\log L(J)\right)^{-2} \\
&\leq \alpha\frac{6}{1+(\pi/\log 2)^2}\sum_{j\geq 0}(1/6 + j\log 2)^{-2} \leq \alpha\frac{6}{1+(\pi/\log 2)^2}\left(1/6 + (\log 2)^{-2}\sum_{j\geq 1}j^{-2}\right) \\
&= \alpha,
\end{aligned}
$$

where the last equation is due to $\sum_{j\geq 1}j^{-2} = \pi^2/6$. Consider II. By Markov's inequality

we obtain

$$II \leq \mathrm{E}_{h_0} \max_{J \in \mathcal{I}_n} \left| \frac{4}{\eta_J'(\alpha) v_J (n-1)} \sum_{i<i'} U_i \mathbb{1}_{M_i^c} U_{i'} \mathbb{1}_{M_{i'}^c} b^K(W_i)' A' A b^K(W_{i'}) \right|$$

$$\leq 4n \, \mathrm{E}_{h_0} |U \mathbb{1}_{\{|U|>M_n\}}| \, \mathrm{E}_{h_0} |U \mathbb{1}_{\{|U|>M_n\}}| \max_{J \in \mathcal{I}_n} \frac{\zeta_J^2 \|(G_b^{-1/2} S G^{-1/2})_l^-\|^2}{\eta_J'(\alpha) v_J}$$

$$\leq 4n M_n^{-6} \left(\mathrm{E}_{h_0}[U^4] \right)^2 \zeta_J^2 \max_{J \in \mathcal{I}_n} \frac{s_J^{-2}}{\eta_J'(\alpha) v_J},$$

where the fourth moment of $U = Y - h_0(X)$ is bounded under Assumption 2(i). From Lemma E.2 we deduce $s_J^{-2} \leq \underline{\sigma}^{-2} v_J$. Thus, using definition $M_n = \sqrt{n} \, \zeta_{\overline{J}}^{-1} (\log \log \overline{J})^{-3/4}$ gives $II = o\left(n^{-2}(\log \log \overline{J})^{9/2} \zeta_{\overline{J}}^8\right) = o(1)$, due to the rate restrictions imposed Assumption 4(i). Consider III. Lemma E.5 implies

$$\mathrm{P}_{h_0}\left(\max_{J \in \mathcal{I}_n} \left| \frac{1}{\eta_J'(\alpha) v_J (n-1)} \sum_{i \neq i'} U_i U_{i'} b^K(W_i)' \left(A'A - \widehat{A}'\widehat{A}\right) b^K(W_{i'}) \right| > \frac{1}{2} \right) = o(1),$$

using $\eta_J'(\alpha) \geq (1 - c_0) \sqrt{\log \log J}/4$ by Lemma E.9 and hence $III = o(1)$.

Step 2: We control the type II error of the test statistic $\widetilde{\mathrm{T}}_n$ where $\eta_J'(\alpha)$ is replaced by $\eta_J''(\alpha) > 0$. From the definition $\overline{J} = \sup\{J : s_J^{-1} \zeta_J^2 \sqrt{(\log J)/n} \leq \overline{c}\}$ we infer that the dimension parameter J_0 is given in (4.3) satisfies $\underline{J} \leq J_0 \leq \overline{J}/2$ for $\overline{c}$ sufficiently large by Assumption 4(iii). Thus, by the construction of the set $\mathcal{I}_n$ there exists $J^* \in \mathcal{I}_n$ such that $J_0 \leq J^* < 2J_0$. We denote $K^* = K(J^*)$. We further evaluate for all $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ that

$$\mathrm{P}_h\left(\widetilde{\mathrm{T}}_n = 0 \right) = \mathrm{P}_h\left(n \, \widehat{D}_J(h_0) \leq \eta_J''(\alpha) \, v_J \text{ for all } J \in \mathcal{I}_n \right)$$

$$\leq \mathrm{P}_h\left(n \, \widehat{D}_{J^*}(h_0) \leq c_1 \sqrt{\log \log n} \, v_{J^*} \right),$$

where in the last inequality we used Lemma E.9 which yields $\eta_{J^*}''(\alpha) \leq c_1 \sqrt{\log \log n}$ with $c_1 = 4(1 + c_0)$. Lemma B.1 implies that $\sup_{h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)} \mathrm{P}_h\left(\widetilde{\mathrm{T}}_n = 0 \right) = o(1)$.

Step 3: Finally, we account for estimation of the normalization factor v_J and for estimation of upper bound of the RES index $\widehat{\mathcal{I}}_n$. We control the type I error of the test $\widehat{\mathrm{T}}_n$ under simple null hypotheses as follows

$$\mathrm{P}_{h_0}\left(\widehat{\mathrm{T}}_n = 1 \right) \leq \mathrm{P}_{h_0}\left(\max_{J \in \mathcal{I}_n} \left\{ n\widehat{D}_J(h_0)/(\eta_J(\alpha)\widehat{v}_J) \right\} > 1, \quad \widehat{v}_J \geq (1 - c_0) v_J \text{ for all } J \in \mathcal{I}_n \right)$$

$$+ \mathrm{P}_{h_0}\left(\widehat{v}_J < (1 - c_0) v_J \text{ for all } J \in \mathcal{I}_n \right) + \mathrm{P}_{h_0}\left(\widehat{J}_{\max} > \overline{J} \right)$$

$$\leq \mathrm{P}_{h_0}\left(\max_{J \in \mathcal{I}_n} \left\{ n\widehat{D}_J(h_0)/(\eta_J(\alpha) v_J) \right\} > 1 - c_0 \right) + \mathrm{P}_{h_0}\left(\max_{J \in \mathcal{I}_n} |\widehat{v}_J/v_J - 1| > c_0 \right) + o(1)$$

$$\leq \mathrm{P}_{h_0}\left(\max_{J \in \mathcal{I}_n} \left\{ n\widehat{D}_J(h_0)/(\eta_J'(\alpha) v_J) \right\} > 1 \right) + o(1) \leq \alpha + o(1),$$

where the second inequality is due to Lemma E.10(i), the third inequality is due to Lemma E.6, and the last inequality is due to Step 1 of this proof. To bound the type II error of the test $\widehat{T}_n$ recall the definition of $J^* \in \mathcal{I}_n$ given in Step 2. Using again Lemmas E.10(ii) and E.6 we evaluate uniformly for $h \in \mathcal{H}_1(\delta^\circ, r_n)$:

$$P_h(\widehat{T}_n = 0) \leq P_h\left(n\widehat{D}_{J^*}(h_0) \leq \eta_{J^*}(\alpha)\,\widehat{v}_{J^*}\right) + P_h\left(J^* > \widehat{J}_{\max}\right)$$

$$\leq P_h\left(n\widehat{D}_{J^*}(h_0) \leq \eta_{J^*}(\alpha)\,\widehat{v}_{J^*}, \quad \widehat{v}_{J^*} \leq (1+c_0)v_{J^*}\right) + P_h\left(\widehat{v}_{J^*} > (1+c_0)v_{J^*}\right) + o(1)$$

$$= P_h\left(n\widehat{D}_{J^*}(h_0) \leq (1+c_0)\eta_{J^*}(\alpha)\,v_{J^*}\right) + o(1) = P_h\left(n\widehat{D}_{J^*}(h_0) \leq \eta_{J^*}''(\alpha)v_{J^*}\right) + o(1) = o(1),$$

where the last equation is due to Step 2 of this proof. $\qquad\square$

Lemma B.1. *(i) Under the conditions of Theorem 3.2 we have for some constant $c_0 > 0$:*

$$\sup_{h \in \mathcal{H}_1(\delta, r_{n,J})} P_h\left(n\widehat{D}_J(h_0) \leq (1+c_0)\eta_J(\alpha)v_J\right) = o(1).$$

(ii) Under the conditions of Theorem 4.1 we have

$$\sup_{h \in \mathcal{H}_1(\delta^\circ, r_n)} P_h\left(n\,\widehat{D}_{J^*}(h_0) \leq c_1\sqrt{\log\log n}\,v_{J^*}\right) = o(1),$$

where J^* and c_1 are given in the proof of Theorem 4.1.

Proof. It is sufficient to prove (ii). We make use of the notation $B_J = (\|\,E_h[V^J]\| - \|h - h_0\|_\mu)^2$. Using the inequality $\|\,E_h[V^{J^*}]\|^2 \geq \|h - h_0\|_\mu^2/2 - B_{J^*}$, we derive

$$P_h\left(n\,\widehat{D}_{J^*} \leq c_1\sqrt{\log\log n}\,v_{J^*}\right) = P_h\left(\|\,E_h[V^{J^*}]\|^2 - \widehat{D}_{J^*} > \|\,E_h[V^{J^*}]\|^2 - \frac{c_1\sqrt{\log\log n}\,v_{J^*}}{n}\right)$$

$$\leq P_h\left(\left|\frac{4}{n(n-1)}\sum_{j=1}^{J^*}\sum_{i<i'}\left(V_{ij}V_{i'j} - E_h[V_{1j}]^2\right)\right| > \rho_{J^*}(h)\right)$$

$$+ P_h\left(\left|\frac{4}{n(n-1)}\sum_{i<i'}(Y_i - h_0(X_i))(Y_{i'} - h_0(X_{i'})b^{K^*}(W_i)'\left(A'A - \widehat{A}'\widehat{A}\right)b^{K^*}(W_{i'})\right| > \rho_{J^*}(h)\right) = T_1 + T_2,$$

where $\rho_{J^*}(h) = \|h - h_0\|_\mu^2/2 - c_1 n^{-1}\sqrt{\log\log n}\,v_{J^*} - B_{J^*}$. We first derive an upper bound for the term B_{J^*}. The definitions of V^J and Q_J imply

$$\|\,E_h[V^{J^*}]\|^2 = E_h[(Y - h_0(X))b^{K^*}(W)']A'A\,E_h[(Y - h_0(X))b^{K^*}(W)]$$

$$= \left\|G^{1/2}(G_b^{-1/2}S)_l^- G_b^{-1/2}\,E[(h - h_0)(X)b^{K^*}(W)]\right\|^2 = \|Q_{J^*}(h - h_0)\|_\mu^2.$$

Consequently, from Lemma E.3 we infer

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

for some constant C_B, due to the definition of J^*. To establish an upper bound of T_1, we make use of inequality (E.3) together with Markov's inequality which yields

$$T_1 = O\left(\frac{n^{-1}\|\langle Q_{J^*}(h-h_0), \psi^{J^*}\rangle'_\mu (G_b^{-1/2}S)_l^-\|^2 + n^{-2}v_{J^*}^2}{\rho_{J^*}^2(h)}\right). \tag{B.2}$$

In the following, we distinguish between two cases. First, consider the case where $n^{-2}v_{J^*}^2$ dominates the summand in the numerator. 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

$$\rho_{J^*}(h) = \|h-h_0\|_\mu^2/2 - c_1 n^{-1}\sqrt{\log\log n}\,v_{J^*} - B_{J^*} \geq (\delta^\circ/2-1-c_0-C_B)\,\mathsf{r}_n^2 = \kappa_0 \mathsf{r}_n^2 \tag{B.3}$$

for some constant $\kappa_0 := \delta^\circ/2 - 1 - c_0 - C_B$ which is positive for any $\delta^\circ > 2(1+c_0+C_B)$. From inequality (B.2) we infer $T_1 \leq O\left(n^{-2}v_{J^*}^2\kappa_0^{-2}\mathsf{r}_n^{-4}\right)$ which becomes arbitrary small for δ° sufficiently large. Second, consider the case where $n^{-1}\|\langle Q_{J^*}(h-h_0), \psi^{J^*}\rangle'_\mu(G_b^{-1/2}S)_l^-\|^2$ dominates. Now using $\|(G_b^{-1/2}SG^{-1/2})_l^-\| = s_{J^*}^{-1}$ together with the notation $\widetilde{\psi}^J = G^{-1/2}\psi^J$ we obtain

$$\frac{1}{n}\|\langle Q_{J^*}(h-h_0), \psi^{J^*}\rangle'_\mu(G_b^{-1/2}S)_l^-\|^2 = \frac{1}{n}\|\langle Q_{J^*}(h-h_0), \psi^{J^*}\rangle'_\mu G^{-1/2}(G_b^{-1/2}SG^{-1/2})_l^-\|^2$$
$$= O\left(n^{-1}s_{J^*}^{-2}\|\langle Q_{J^*}(h-h_0), \widetilde{\psi}^{J^*}\rangle_\mu\|^2\right) = O\left(n^{-1}s_{J^*}^{-2}(\|h-h_0\|_\mu^2 + (J^*)^{-2p/d_x})\right), \tag{B.4}$$

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

$$\rho_{J^*}(h) = \frac{\|h-h_0\|_\mu^2}{2} - \frac{c_1\sqrt{\log\log n}\,v_{J^*}}{n} + B_{J^*} \geq \left(\frac{1}{2} - \frac{1+c_0}{\delta^\circ} - \frac{C_B}{\delta^\circ}\right)\|h-h_0\|_\mu^2 = \kappa_1\|h-h_0\|_\mu^2$$

for some constant $\kappa_1 := 1/2 - (1+c_0+C_B)/\delta^\circ$ which is positive for any $\delta^\circ > 2(1+c_0+C_B)$. Hence, inequality (B.2) yields uniformly for $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ that

$$T_1 = O\left(n^{-1}s_{J^*}^{-2}\left(\frac{1}{\|h-h_0\|_\mu^2} + \frac{1}{\|h-h_0\|_\mu^4(J^*)^{2p/d}}\right)\right) = O\left(n^{-1}s_{J^*}^{-2}\mathsf{r}_n^{-2}\right) = o(1).$$

Finally, $T_2 = o(1)$ uniformly for $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ by making use of Lemma E.4. $\qquad\square$

Proof of Corollary 4.1. We show that Assumption 4(ii) is automatically satisfied in the mildly and severely ill-posed cases. From Lemmas E.1 and E.2 we infer $v_J \leq \overline{\sigma}^2 s_J^{-2}\sqrt{J}$ and $v_L \geq \underline{\sigma}^2\left(\sum_{l=1}^L s_l^{-4}\right)^{1/2}$. In the mildly ill-posed case ($\tau_j \sim j^{a/d_x}$), we obtain

$$\frac{v_J}{v_L} \lesssim \frac{J^{2a/d_x+1/2}}{\sqrt{\sum_{l=1}^L l^{4a/d_x}}} \lesssim \frac{J^{2a/d_x+1/2}}{L^{2a/d_x+1/2}-1} = o(1)$$

for all $J = o(L)$. In the severely ill-posed case ($\tau_j \sim \exp(j^{a/d_x}/2)$), we evaluate

$$\frac{v_J}{v_L} \lesssim \frac{\exp(J^{a/d_x} + \log(J)/2)}{\exp(L^{a/d_x})} \lesssim \frac{\exp(J^{a/d_x} + \log(J)/2)}{\exp(2J^{a/d_x})} = \frac{\sqrt{J}}{\exp(J^{a/d_x})} = o(1)$$

since $J = o(L)$ and J diverges as the sample size n tends to infinity.

We may now apply Theorem 4.1 which establishes the rate $r_n = J_0^{-p/d_x}$ where $J_0 = \max\left\{ J : n^{-1}\sqrt{\log\log n}\, s_J^{-2}\sqrt{J} \leq J^{-2p/d_x} \right\}$. In the mildly ill-posed case, we obtain $J_0 \sim \left(n/\sqrt{\log\log n}\right)^{2d_x/(4(p+a)+d_x)}$ which implies $r_n = \left(\sqrt{\log\log n}/n\right)^{2p/(4(p+a)+d_x)}$. In the severely ill-posed case, note that if $J_0 \sim \left(c\log n\right)^{d_x/a}$ for some constant $c \in (0,1)$ then we obtain $n^{-1}\sqrt{\log\log n}\, s_{J_0}^{-2}\sqrt{J_0} \lesssim J_0^{-2p/d_x} \sim \left(\log n\right)^{-2p/d_x}$. $\square$

Proof of Theorem 4.2. We prove this result in three steps. First, we bound the type I error of the test statistic

$$\widetilde{T}_n = \mathbb{1}\left\{ \max_{J \in \mathcal{I}_n}\left\{ n\widehat{D}_J/(\widehat{\eta}_J(\alpha)v_J)\right\} > 1 \right\}.$$

Second, we bound the type II error of $\widetilde{T}_n$. Third, we show that Steps 1 and 2 are sufficient to control the first and second type errors of our adaptive test statistic $\widehat{T}_n$ for the composite null.

Step 1: We control the type I error of the test statistic $\widetilde{T}_n$ using the decomposition

$$n(n-1)\,\widehat{D}_J = \sum_{i \neq i'} \left(Y_i - \widehat{h}_J^{\mathrm{R}}(X_i)\right)\left(Y_{i'} - \widehat{h}_J^{\mathrm{R}}(X_{i'})\right)b^K(W_i)'\widehat{A}'\widehat{A}b^K(W_{i'})$$

$$= \left\| \sum_i \left(Y_i - \widehat{h}_J^{\mathrm{R}}(X_i)\right)\widehat{A}b^K(W_i)\right\|^2 - \sum_i \left\|\left(Y_i - \widehat{h}_J^{\mathrm{R}}(X_i)\right)\widehat{A}b^K(W_i)\right\|^2.$$

From the definition of the restricted sieve NPIV estimator in (2.4) we infer:

$$\left\| n^{-1}\sum_i \left(Y_i - \widehat{h}_J^{\mathrm{R}}(X_i)\right)\widehat{A}b^K(W_i)\right\| = \left(n^{-1}\sum_i \left(\widehat{h}_J(X_i) - \widehat{h}_J^{\mathrm{R}}(X_i)\right)^2 \mu(X_i)\right)^{1/2}$$

$$\leq \left(n^{-1}\sum_i \left(\widehat{h}_J(X_i) - \Pi_J h(X_i)\right)^2 \mu(X_i)\right)^{1/2} = \left\| n^{-1}\sum_i \left(Y_i - h(X_i)\right)\widehat{A}b^K(W_i)\right\| + o_p(1)$$

uniformly in $h \in \mathcal{H}_0$, where the last equation holds uniformly in $J \in \mathcal{I}_n$ using that $\max_{J \in \mathcal{I}_n}\|\Pi_J^{\mathcal{I}_n}h - h\|_\mu = o(1)$ (since $\underline{J}$ goes to infinity and $\|h - \Pi_{\underline{J}}h\|_\mu = o(1)$). Con-

sequently, uniformly in $h \in \mathcal{H}_0$ and $J \in \mathcal{I}_n$ we have

$$
n\widehat{D}_J \leq n\widehat{D}_J(h) + \frac{1}{n}\sum_i \left\|\left(\widehat{h}_J^{\mathrm{R}} - h\right)(X_i)\widehat{A}b^K(W_i)\right\|^2
$$

$$
+ 2\frac{1}{n}\sum_i \left(Y_i - h(X_i)\right)b^K(W_i)'\widehat{A}'\widehat{A}b^K(W_i)\left(\widehat{h}_J^{\mathrm{R}} - h\right)(X_i) + o_p(1)
$$

$$
= n\widehat{D}_J(\Pi_{\mathrm{R}}h) + T_{1,J} + 2T_{2,J} + o_p(1).
$$

Lemma E.9 together with Assumption 5(iii), i.e., $\inf_{h \in \mathcal{H}_0} \mathrm{P}_h\left(\forall J \in \mathcal{I}_n : J^c \leq \widehat{\gamma}_J\right) = 1 + o(1)$ for some $0 < c \leq 1$, implies $\inf_{h \in \mathcal{H}_0} \mathrm{P}_h\left(\forall J \in \mathcal{I}_n : \sqrt{\log\log(J) + \log(c/\alpha)} \leq 4\widehat{\eta}_J(\alpha)\right) = 1 + o(1)$. In particular, this gives $\inf_{h \in \mathcal{H}_0} \mathrm{P}_h\left(\forall J \in \mathcal{I}_n : \sqrt{\log\log(J)} \leq 8\widehat{\eta}_J(\alpha)\right) = 1 + o(1)$. Now we may follow Step 1 of the proof of Theorem 4.1 and obtain

$$
\limsup_{n \to \infty} \sup_{h \in \mathcal{H}_0} \mathrm{P}_h\left(\max_{J \in \mathcal{I}_n}\left\{n\widehat{D}_J(\Pi_{\mathrm{R}}h)/(\widehat{\eta}_J(\alpha)\mathrm{v}_J)\right\} > 1\right) \leq \alpha.
$$

It remains to control $T_{1,J}$ and $T_{2,J}$. Consider $T_{1,J}$. We observe

$$
T_{1,J} = \frac{1}{n}\sum_i \left\|\left(\widehat{h}_J^{\mathrm{R}} - h\right)(X_i)Ab^K(W_i)\right\|^2 + \frac{1}{n}\sum_i \left\|\left(\widehat{h}_J^{\mathrm{R}} - h\right)(X_i)\left(\widehat{A} - A\right)b^K(W_i)\right\|^2
$$

$$
= T_{11,J} + T_{12,J}.
$$

Using the notation $\widetilde{b}^K(\cdot) = G_b^{-1/2}b^K(\cdot)$, we evaluate

$$
\max_{J \in \mathcal{I}_n} \frac{T_{11,J}}{\mathrm{v}_J\sqrt{\log\log J}} \leq \max_{J \in \mathcal{I}_n} \frac{\left(\|\widehat{h}_J^{\mathrm{R}} - h\|_\mu \sup_w \|\widetilde{b}^K(w)\|\left\|\left(G_b^{-1/2}SG^{-1/2}\right)_l^-\right\|\right)^2}{\mathrm{v}_J\sqrt{\log\log J}} + o_p(1)
$$

$$
\leq \max_{J \in \mathcal{I}_n} \frac{\left(\|\widehat{h}_J^{\mathrm{R}} - h\|_\mu \zeta_J s_J^{-1}\right)^2}{\mathrm{v}_J\sqrt{\log\log J}} + o_p(1) = o_p(1)
$$

uniformly for $h \in \mathcal{H}_0$, where the last inequality is due to Lemma E.2, i.e., $s_J^{-2} \leq \underline{\sigma}^{-2}\mathrm{v}_J$, and the last equation follows from Assumption 5(i), i.e., $\max_{J \in \mathcal{I}_n} \|\widehat{h}_J^{\mathrm{R}} - h\|_\mu \zeta_J/(\log\log J)^{1/4} = o_p(1)$ uniformly for $h \in \mathcal{H}_0$. Similarly, we obtain $\max_{J \in \mathcal{I}_n} T_{12,J}/(\mathrm{v}_J\sqrt{\log\log J}) = o_p(1)$ uniformly for $h \in \mathcal{H}_0$, using that uniformly for $J \in \mathcal{I}_n$ we have

$$
\left\|\left(\widehat{A} - A\right)G_b^{1/2}\right\| = \left\|\left(\widehat{G}_b^{-1/2}\widehat{S}\widehat{G}^{-1/2}\right)_l^- \widehat{G}_b^{-1/2}G_b^{1/2} - \left(G_b^{-1/2}SG^{-1/2}\right)_l^-\right\|
$$

$$
= O_p(1) \times \left\|\left(\widehat{G}_b^{-1/2}\widehat{S}\right)_l^- \widehat{G}_b^{-1/2}G_b^{1/2} - \left(G_b^{-1/2}S\right)_l^-\right\| = O_p\left(s_J^{-2}\zeta_J\sqrt{(\log J)/n}\right), \quad \text{(B.5)}
$$

where the last bound is due to Chen et al. [2021, Lemma C.4(i)] (with τ_J replaced by s_J^{-1}).

Consider $T_{2,J}$. For all $J \in \mathcal{I}_n$ we evaluate

$$T_{2,J} \leq \frac{1}{n} \sum_i \left(Y_i - h(X_i)\right) b^K(W_i)' A' A b^K(W_i) \left(\widehat{h}_J^{\mathrm{R}} - h\right)(X_i)$$
$$+ \frac{1}{n} \sum_i \left(Y_i - h(X_i)\right) b^K(W_i)' \left(\widehat{A}' \widehat{A} - A'A\right) b^K(W_i) \left(\widehat{h}_J^{\mathrm{R}} - h\right)(X_i) = T_{21,J} + T_{22,J}.$$

Consider $T_{21,J}$. We first observe by the Cauchy–Schwarz inequality that

$$T_{21,J} \leq \left(\frac{1}{n} \sum_i \left(Y_i - h(X_i)\right)^2 \|A b^K(W_i)\|^4\right)^{1/2} \left(\frac{1}{n} \sum_i \left(\widehat{h}_J^{\mathrm{R}}(X_i) - h(X_i)\right)^2\right)^{1/2}.$$

Further, the moment bounds imposed in Assumptions 1(i) and 2(i) imply

$$\mathrm{E}\left[\left(Y - h(X)\right)^2 \|A b^K(W)\|^4\right] \leq \overline{\sigma}^2 \zeta_J^2 \mathrm{E} \|G_b^{1/2} A' A b^K(W)\|^2 \leq \overline{\sigma}^2 \zeta_J^2 \|A G_b A'\|_F^2$$
$$\leq \overline{\sigma}^2 \underline{\sigma}^{-2} \zeta_J^2 \|A G_b^{-1/2} \mathrm{E}\left[\left(Y - h(X)\right)^2 b^K(W) b^K(W)'\right] G_b^{-1/2} A'\|_F^2 = \overline{\sigma}^2 \underline{\sigma}^{-2} \zeta_J^2 v_J^2,$$

where the last equation follows from the definition of the normalization term v_J. Consequently, we evaluate

$$\max_{J \in \mathcal{I}_n} \frac{T_{21,J}}{v_J \sqrt{\log \log J}} = \max_{J \in \mathcal{I}_n} \frac{\zeta_J \|\widehat{h}_J^{\mathrm{R}} - h\|_\mu}{\sqrt{\log \log J}} \times O_p \left(\max_{J \in \mathcal{I}_n} \frac{\mathrm{E}[(Y - h(X))^2 \|A b^K(W)\|^4]}{\zeta_J v_J}\right)$$
$$= o_p(1)$$

uniformly for $h \in \mathcal{H}_0$, where the last equation follows from the rate condition imposed in Assumption 5(i). Similarly as above, we obtain that $\max_{J \in \mathcal{I}_n} T_{22,J}/(v_J \sqrt{\log \log J}) = o_p(1)$ uniformly for $h \in \mathcal{H}_0$, using again the upper bound (B.5).

Step 2: We control the second type error of the test statistic $\widetilde{\mathrm{T}}_n$. Let J^* be as in the proof of Theorem 4.1. We obtain for all $h \in \mathcal{H}_1(\delta^\circ, r_n)$ and for some constant $c_2 > 0$:

$$\mathrm{P}_h\left(\widetilde{\mathrm{T}}_n = 0\right) \leq \mathrm{P}_h\left(n \widehat{D}_{J^*}(\widehat{h}_{J\bullet}^{\mathrm{R}}) - c_2 r_n^2 \leq \widehat{\eta}_{J^*}(\alpha) \, v_{J^*}\right) + \mathrm{P}_h\left(\widehat{D}_{J^*}(\widehat{h}_{J\bullet}^{\mathrm{R}}) - c_2 r_n^2 > \widehat{D}_{J^*}\right)$$
$$\leq \mathrm{P}_h\left(\left|\|\mathrm{E}_h[V^{J^*}]\|^2 - \widehat{D}_{J^*}(\Pi_{\mathrm{R}} h)\right| > \|\mathrm{E}_h[V^{J^*}]\|^2/2 - c_2 r_n^2 - \frac{\widehat{\eta}_{J^*}(\alpha) \, v_{J^*}}{2n}\right)$$
$$+ \mathrm{P}_h\left(\left|\widehat{D}_{J^*}(\widehat{h}_{J\bullet}^{\mathrm{R}}) - \widehat{D}_{J^*}(\Pi_{\mathrm{R}} h)\right| > \|\mathrm{E}_h[V^{J^*}]\|^2/2 - c_2 r_n^2 - \frac{\widehat{\eta}_{J^*}(\alpha) \, v_{J^*}}{2n}\right)$$
$$+ \mathrm{P}_h\left(\left|\widehat{D}_{J^*} - \widehat{D}_{J^*}(\widehat{h}_{J\bullet}^{\mathrm{R}})\right| > c_2 r_n^2\right),$$

where the first summand on the right hand side tends to zero, following Step 2 in proof of Theorem 4.1. Consider the second summand on the right hand side. The definition of the

43

estimator $\widehat{D}_J(\Pi_{\mathrm{R}}h)$ implies for all $J \in \mathcal{I}_n$ and $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ that

$$
\widehat{D}_J(\widehat{h}^{\mathrm{R}}_{J\bullet}) - \widehat{D}_J(\Pi_{\mathrm{R}}h) = \frac{1}{n(n-1)} \sum_{i \neq i'} (\widehat{h}^{\mathrm{R}}_{J\bullet} - \Pi_{\mathrm{R}}h)(X_i)(\widehat{h}^{\mathrm{R}}_{J\bullet} - \Pi_{\mathrm{R}}h)(X_{i'})b^K(W_i)'A'Ab^K(W_{i'})
$$

$$
+ \frac{1}{n(n-1)} \sum_{i \neq i'} (\widehat{h}^{\mathrm{R}}_{J\bullet} - \Pi_{\mathrm{R}}h)(X_i)(\widehat{h}^{\mathrm{R}}_{J\bullet} - \Pi_{\mathrm{R}}h)(X_{i'})b^K(W_i)'\Big(A'A - \widehat{A}'\widehat{A}\Big)b^K(W_{i'})
$$

$$
+ \frac{2}{n(n-1)} \sum_{i \neq i'} \big(Y_i - \Pi_{\mathrm{R}}h(X_i)\big)(\widehat{h}^{\mathrm{R}}_{J\bullet} - \Pi_{\mathrm{R}}h)(X_{i'})b^K(W_i)'A'Ab^K(W_{i'})
$$

$$
+ \frac{2}{n(n-1)} \sum_{i \neq i'} \big(Y_i - \Pi_{\mathrm{R}}h(X_i)\big)(\widehat{h}^{\mathrm{R}}_{J\bullet} - \Pi_{\mathrm{R}}h)(X_{i'})b^K(W_i)'\Big(A'A - \widehat{A}'\widehat{A}\Big)b^K(W_{i'})
$$

$$
= T_{3,J} + T_{4,J} + T_{5,J} + T_{6,J}.
$$

Consider $T_{3,J}$. From the proof of Lemma B.1(ii) (see lower bound (B.3)) we obtain

$$
\| \operatorname{E}_h[V^{J^*}] \|^2/2 - c_2\mathsf{r}_n^2 - \sqrt{\log\log(J^*)}\, v_{J^*}/(2n) \geq \kappa_0\mathsf{r}_n^2
$$

for some constant $\kappa_0 := \delta^\circ/2 - c_2 - 1 - c_0 - C_B$ which is positive for δ° sufficiently large. Below, Π^{R}_J denotes the projection onto $\mathcal{H}^{\mathrm{R}}_J = \Psi_J \cap \mathcal{H}^{\mathrm{R}}$. It is thus sufficient to consider

$$
\mathrm{P}_h\Big(T_{3,J^*} > \kappa_0\mathsf{r}_n^2\Big) \leq \mathrm{P}_h\Big(\Big\|A\frac{1}{n}\sum_i (\widehat{h}^{\mathrm{R}}_{J\bullet} - \Pi_{\mathrm{R}}h)(X_i)b^{K^*}(W_i)\Big\|^2 > \kappa_0\mathsf{r}_n^2\Big)
$$

$$
\leq \mathrm{P}_h\Big(\|\widehat{h}^{\mathrm{R}}_{J\bullet} - \Pi^{\mathrm{R}}_{J\bullet}h\|_\mu^2\|A\widehat{S}G^{-1/2}\|^2 + \Big\|A\frac{1}{n}\sum_i (\Pi^{\mathrm{R}}_{J\bullet}h - \Pi_{\mathrm{R}}h)(X_i)b^{K^*}(W_i)\Big\|^2 > \kappa_0\mathsf{r}_n^2\Big)
$$

$$
\leq \mathrm{P}_h\Big(\|\widehat{h}^{\mathrm{R}}_{J\bullet} - \Pi^{\mathrm{R}}_{J\bullet}h\|_\mu^2 + \Big\|A\operatorname{E}\big[(\Pi^{\mathrm{R}}_{J\bullet}h - \Pi_{\mathrm{R}}h)(X)b^{K^*}(W)\big]\Big\|^2 > \kappa_0\mathsf{r}_n^2\Big) + o(1)
$$

uniformly for $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$. From Lemma E.3 we infer

$$
\Big\|A\operatorname{E}\big[(\Pi^{\mathrm{R}}_{J\bullet}h - \Pi_{\mathrm{R}}h)(X)b^{K^*}(W)\big]\Big\|^2 = \|Q_{J^*}(\Pi^{\mathrm{R}}_{J\bullet}h - \Pi_{\mathrm{R}}h)\|_\mu^2 \leq C_B\, \mathsf{r}_n^2
$$

by Assumption 5(ii). Consequently, we obtain

$$
\sup_{h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)} \mathrm{P}_h\Big(T_{3,J^*} > \kappa_0\mathsf{r}_n^2\Big) \leq \sup_{h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)} \mathrm{P}_h\Big(\|\widehat{h}^{\mathrm{R}}_{J\bullet} - \Pi^{\mathrm{R}}_{J\bullet}h\|_\mu^2 > (\kappa_0 - C_B)\mathsf{r}_n^2\Big) + o(1) \leq o(1)
$$

for δ° sufficiently large, by Assumption 5(ii). Next we consider $T_{4,J}$. We evaluate

$$
|T_{4,J^*}| \leq \Big\|\frac{1}{n}\sum_i (\widehat{h}^{\mathrm{R}}_{J\bullet} - \Pi_{\mathrm{R}}h)(X_i)\widetilde{b}^{K^*}(W_i)\Big\|^2\|(\widehat{A} - A)G_b^{1/2}\|^2.
$$

Chen and Christensen [2018, Lemma F.10(a)] yields $\|(\widehat{A}-A)G_b^{1/2}\|^2 = O_p(n^{-1}s_{J^*}^{-4}\zeta_{J^*}^2(\log J^*))$ and hence, we obtain for δ° sufficiently large that $\sup_{h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)}\mathrm{P}_h(T_{4,J^*} > \kappa_0\mathsf{r}_n^2) = o(1)$, due to Assumption 4(ii). The bounds on T_{5,J^*} and T_{6,J^*} follow analogously. It remains

to control $P_h\big(|\widehat{D}_{J^*} - \widehat{D}_{J^*}(\widehat{h}_{J\bullet}^{\mathrm{R}})| > c_2 r_n^2\big)$. We have $\|\widehat{h}_{J\bullet}^{\mathrm{R}} - \widehat{h}_{J^*}^{\mathrm{R}}\|_\mu \le c\, r_n$, for some constant $c > 0$, by using that $\|\Pi_{J\bullet}\widehat{h}_{J^*}^{\mathrm{R}} - \widehat{h}_{J^*}^{\mathrm{R}}\|_\mu \le \sup_{h\in\mathcal{H}^{\mathrm{R}}} \|\Pi_{J\bullet}h - h\|_\mu = O(r_n)$ due to Assumption 5(ii). Consequently, by following the derivations of the upper bounds of T_{3,J^*} and T_{4,J^*} we obtain

$$\sup_{h\in\mathcal{H}_1(\delta^\circ, r_n)} P_h\left(\big|\widehat{D}_{J^*} - \widehat{D}_{J^*}(\widehat{h}_{J\bullet}^{\mathrm{R}})\big| > c_2 r_n^2\right) = o(1),$$

using that c_2 can be chosen sufficiently large (depending on δ°).

Step 3: Finally, we account for estimation of the normalization factor v_J and for estimation of the upper bound of the RES index set $\widehat{\mathcal{I}}_n$. Lemma E.10(i) implies $\sup_{h\in\mathcal{H}_0} P_h\big(\widehat{J}_{\max} > \overline{J}\big) = o(1)$. We thus control the type I error of the test $\widehat{\mathsf{T}}_n$ for testing composite hypotheses, as follows:

$$P_h\left(\widehat{\mathsf{T}}_n = 1\right) \le P_h\left(\max_{J\in\mathcal{I}_n} \frac{n\widehat{D}_J}{\widehat{\eta}_J(\alpha)v_J} > 1 - c_0\right) + P_h\left(\max_{J\in\mathcal{I}_n} |\widehat{v}_J/v_J - 1| > c_0\right) + o(1) \le \alpha + o(1)$$

uniformly for $h \in \mathcal{H}_0$, where the last inequality is due to Step 1 of this proof and Lemma E.6(ii). To bound the type II error of the test $\widehat{\mathsf{T}}_n$ recall the definition of $J^* \in \mathcal{I}_n$ introduced in Step 2. Another application of Lemma E.10(ii) implies uniformly for $h \in \mathcal{H}_1(\delta^\circ, r_n)$:

$$P_h\left(\widehat{\mathsf{T}}_n = 0\right) \le P_h\left(n\widehat{D}_{J^*} \le c\,\widehat{\eta}_{J^*}(\alpha)v_{J^*}\right) + P_h\left(|\widehat{v}_{J^*}/v_{J^*} - 1| > c_0\right) = o(1),$$

where the last equation is due to Step 2 and Lemma E.6(i). $\qquad\square$

Supplement to "Adaptive, Rate-Optimal Hypothesis Testing in Nonparametric IV Models"

CHRISTOPH BREUNIG   XIAOHONG CHEN

First version: August 2018, Revised December 22, 2021

This supplementary appendix contains materials to support our main paper. Appendix C presents additional simulation results. Appendix D provides proofs of our results on confidence sets in Subsection 4.3. Appendix E establishes several technical results. In particular, it provides an upper bound for quadratic distance estimation, which is essential for our upper bound on the minimax rate of testing in L^2. Finally, Appendix F gathers an exponential inequality for U-statistics.

C. Additional Simulations

This section provides additional simulation results. All the simulation results are based on 5000 Monte Carlo replications for every experiment. Due to the lack of space we report simulation results for the nominal level $\alpha = 0.05$ unless stated otherwise.

C.1. Empirical Size and Power for Section 5 with Larger Sample Sizes

Table G below provides additional empirical size results of our adaptive test of monotonicity for the simulation design stated in Subsection 5.1. It replicates Table 1 using a larger sample size $n = 10000$, which indicates that our adaptive test does control size asymptotically.

Figure C below provides additional power comparison for the simulation design stated in Subsection 5.2. It replicates Figure 2 using a larger sample size $n = 1000$. It shows that the power of our adaptive test increases fast as sample size increases.

Table H below provides additional empirical size results of our adaptive test of parametric form for the simulation design stated in Subsection 5.2. It replicates Table 2 using a larger sample size $n = 10000$ for our test $\widehat{\mathrm{T}}_n$ and the t-test. It indicates that our adaptive test still controls size for larger sample sizes.

n	c_0	ξ	$\widehat{\mathrm{T}}_n$ with $K=2J$			$\widehat{J}$	$\widehat{\mathrm{T}}_n$ with $K=4J$			$\widehat{J}$
			10%	5%	1%	at 5%	10%	5%	1%	at 5%
10000	0.01	0.3	0.042	0.015	0.002	3.47	0.051	0.022	0.002	3.43
		0.5	0.056	0.025	0.005	3.84	0.066	0.028	0.006	3.95
		0.7	0.065	0.031	0.007	4.19	0.065	0.033	0.009	4.18
	0.1	0.3	0.028	0.010	0.000	3.52	0.038	0.015	0.001	3.46
		0.5	0.033	0.015	0.003	4.03	0.039	0.020	0.003	4.13
		0.7	0.045	0.022	0.006	4.52	0.044	0.026	0.007	4.51
	1	0.3	0.014	0.005	0.000	3.60	0.018	0.006	0.000	3.50
		0.5	0.013	0.006	0.000	4.27	0.017	0.005	0.001	4.36
		0.7	0.010	0.005	0.001	4.86	0.011	0.004	0.000	4.87

Table G: Testing Monotonicity – Empirical size of our adaptive test $\widehat{\mathrm{T}}_n$. Design from Section 5.1. Replication of Table 1 for our test $\widehat{\mathrm{T}}_n$ with $n = 10000$.
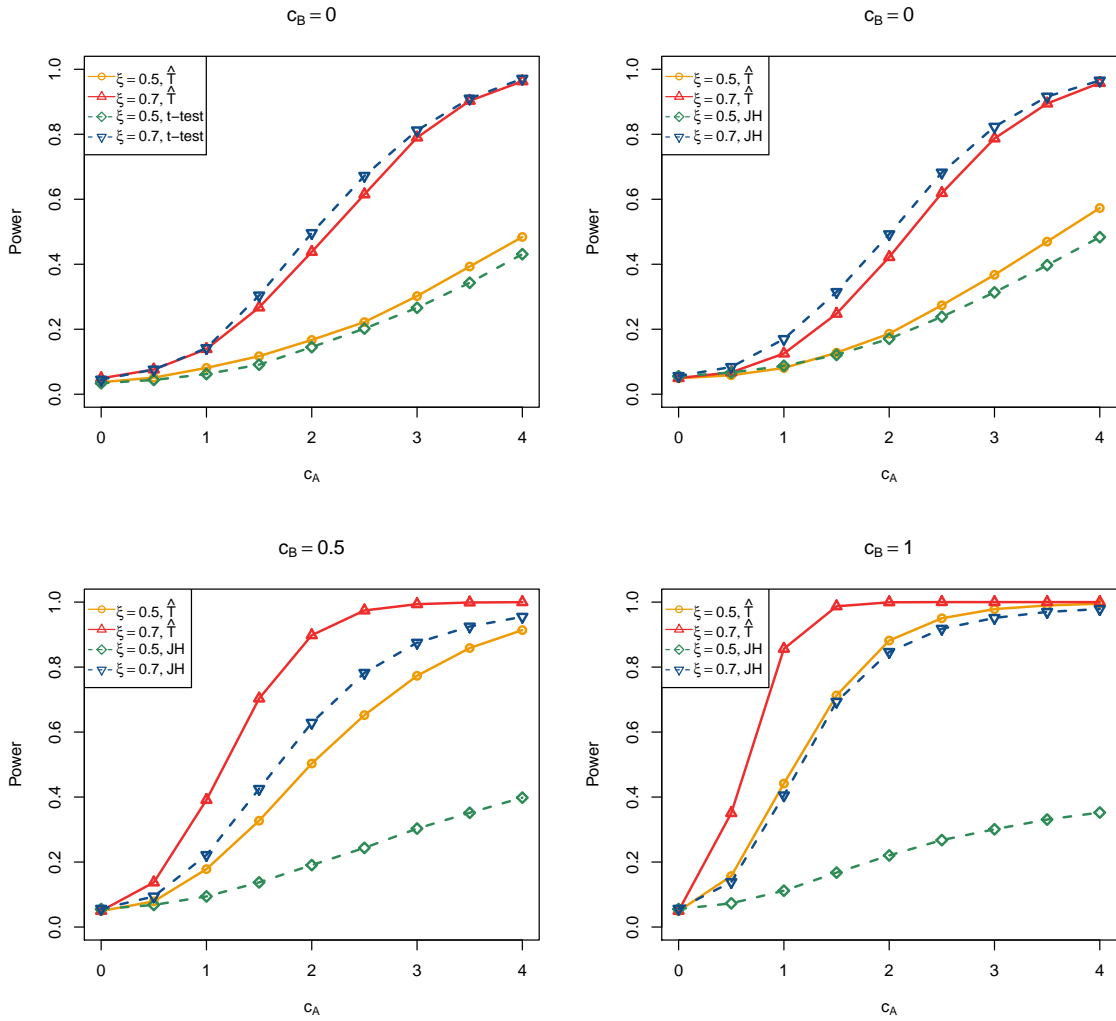

Figure C: Testing Parametric Form - Size-adjusted empirical power of our adaptive test $\widehat{\mathrm{T}}_n$ (solid lines), $\xi = \{0.5, 0.7\}$. First row: power comparison to the t-test and JH test when $c_B = 0$; Second row: power comparison to the JH test when $c_B = 0.5$ and $c_B = 1$. Design from Section 5.2. Replication of Figure 2 with $n = 1000$.

n	ξ	$\widehat{\mathtt{T}}_n$ with $K=2J$	$\widehat{J}$	$\widehat{\mathtt{T}}_n$ with $K=4J$	$\widehat{J}$	t-test
10000	0.3	0.030	3.49	0.035	3.45	0.036
	0.5	0.042	3.85	0.051	3.97	0.047
	0.7	0.055	4.18	0.055	4.17	0.048

Table H: Testing Parametric Form – Empirical size of our adaptive test $\widehat{\mathtt{T}}_n$. 5% nominal level. Design from Section 5.2. Replication of Table 2 for our test $\widehat{\mathtt{T}}_n$ and the t-test with $n = 10000$.

C.2. Adaptive Testing for Monotonicity: Simulation Design II

We generate the dependent variable Y according to the NPIV model (2.1), where

$$h(x) = x/5 + x^2 + c_A \sin(2\pi x) \tag{C.1}$$

and (W^*, ϵ, ν) follows a multivariate standard normal distribution. We set $W = \Phi(W^*)$, $X = \Phi\big(\xi W^* + \sqrt{1-\xi^2}\epsilon\big)$, and $U = (0.3\epsilon + \sqrt{1-(0.3)^2}\nu)/2$. The experimental design with $c_A = 0$ and $\xi \in \{0.3, 0.5\}$ coincides with the one considered by Chetverikov and Wilhelm [2017].

n	c_A	ξ	$\widehat{\mathtt{T}}_n$ with $K=2J$	$\widehat{J}$	$\widehat{\mathtt{T}}_n$ with $K=4J$	$\widehat{J}$	$\mathtt{T}^{\mathtt{B}}_{n,3}$ with $K=4J=12$
500	0.0	0.3	0.001	3.00	0.003	3.02	0.005
		0.5	0.004	3.40	0.004	3.38	0.009
		0.7	0.002	3.75	0.002	3.72	0.004
	0.1	0.3	0.001	3.00	0.005	3.03	0.006
		0.5	0.008	3.39	0.008	3.38	0.009
		0.7	0.007	3.69	0.008	3.65	0.006
1000	0.0	0.3	0.003	3.02	0.005	3.06	0.008
		0.5	0.004	3.67	0.004	3.50	0.009
		0.7	0.003	4.24	0.002	4.32	0.004
	0.1	0.3	0.004	3.02	0.007	3.06	0.010
		0.5	0.007	3.62	0.007	3.48	0.012
		0.7	0.007	4.12	0.005	4.18	0.006
5000	0.0	0.3	0.006	3.45	0.005	3.36	0.011
		0.5	0.003	3.84	0.003	3.90	0.008
		0.7	0.001	4.75	0.001	4.73	0.002
	0.1	0.3	0.009	3.44	0.007	3.35	0.012
		0.5	0.010	3.73	0.009	3.78	0.013
		0.7	0.005	4.53	0.004	4.50	0.005

Table I: Testing Monotonicity - Empirical Size of our adaptive test $\widehat{\mathtt{T}}_n$ and of the nonadaptive bootstrap test $\mathtt{T}^B_{n,3}$. Nominal level $\alpha = 0.05$. Design from Appendix C.2.

We consider the null hypothesis of weakly increasing function h. The null hypothesis is satisfied for h given in (C.1) for $c_A \in [0, 0.184)$, and is violated when $c_A \geq 0.184$. Note that the degree of nonlinearity of h given in (C.1) becomes larger as the constant c_A increases to 1. Table I reports the empirical size of our adaptive test $\widehat{\mathtt{T}}_n$ given in (2.9) in the main paper, with the 5% nominal level, using quadratic B-spline basis functions with varying number

of knots for h. We report simulation results for cases $K = 2J$ and $K = 4J$. In addition, we report the empirical size of the nonadaptive bootstrap test $\mathtt{T}^{\mathtt{B}}_{n,3}$ for monotonicity with $J = 3$ and $K = 4J = 12$. Again we observe that our adaptive tests $\widehat{\mathtt{T}}_n$ and the nonadaptive bootstrap test $\mathtt{T}^{\mathtt{B}}_{n,3}$ provide adequate size control across different design specifications.
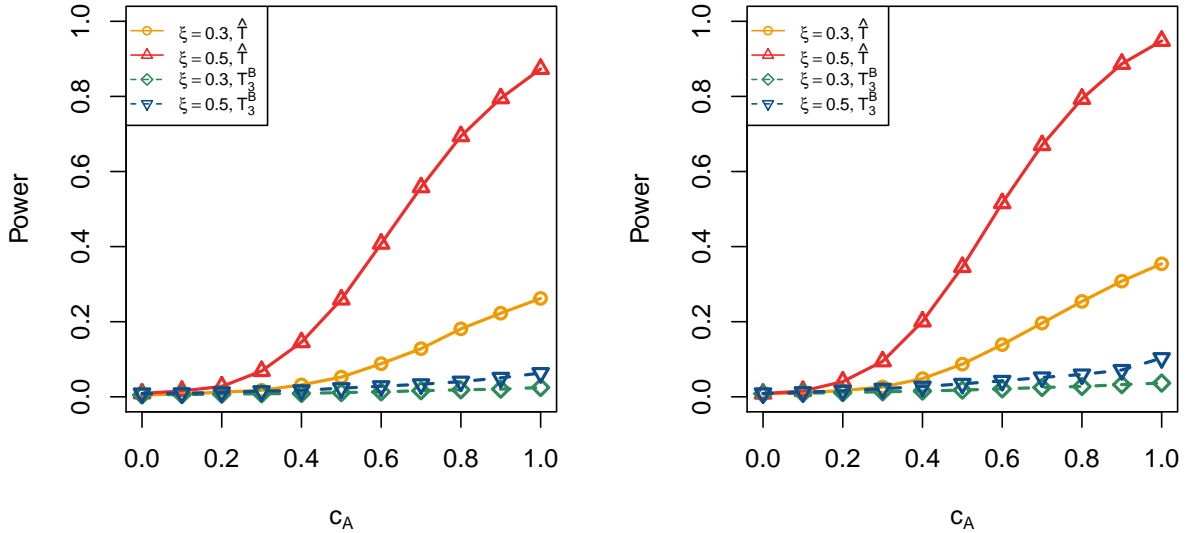

Figure D: Testing Monotonicity – Size-adjusted empirical power of our adaptive $\widehat{\mathtt{T}}_n$ (solid lines) and of the nonadaptive bootstrap test $\mathtt{T}^{\mathtt{B}}_{n,3}$ (dashed lines), $\xi = \{0.3, 0.5\}$. LHS: $n = 500$; RHS: $n = 1000$. Design from Appendix C.2.

Figure D provides empirical rejection probabilities of our adaptive test $\widehat{\mathtt{T}}_n$ and of the nonadaptive bootstrap test $\mathtt{T}^{\mathtt{B}}_{n,3}$ (with a fixed sieve dimension $J = 3$), with $K(J) = 4J$. For all $c_A \geq 0.2$ (the null hypothesis is violated), the nonadaptive bootstrap test $\mathtt{T}^{\mathtt{B}}_{n,3}$ has almost trivial power. In contrast, our adaptive test $\widehat{\mathtt{T}}_n$ has non-trivial power for all $c_A > 0.2$ and its finite sample power increases as $c_A > 0.2$ becomes larger. We see that the substantial improvement in finite- sample power through adaptation even for small sample size $n = 500$.

C.3. Simulations for Multivariate Instruments

This section presents additional simulations for testing parametric hypotheses in the presence of multivariate conditioning variable $W = (W_1, W_2)$. We set $X_i = \Phi(X_i^*)$, $W_{1i} =$

$\Phi(W_{1i}^*)$, and $W_{2i} = \Phi(W_{2i}^*)$, where

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

We generate the dependent variable Y according to the NPIV model (2.1) where

$$
h(x) = -x/5 + c_A x^2. \tag{C.3}
$$

We test the null hypothesis of linearity, i.e., whether $c_A = 0$.

n	Design	ξ	$\widehat{\mathrm{T}}_n$ with $K = 4J$	$\widehat{J}$	$\widehat{\mathrm{IT}}_n$	$\widehat{K}$
500	(5.1)	0.3	0.023	3.12	0.051	4.44
	$d_x = d_w$	0.5	0.030	3.46	0.050	4.44
		0.7	0.032	3.87	0.051	4.42
	(C.2)	0.3	0.035	3.46	0.038	8.99
	$d_x < d_w$	0.5	0.039	3.49	0.042	8.97
		0.7	0.039	3.88	0.037	8.89
1000	(5.1)	0.3	0.023	3.17	0.045	4.40
		0.5	0.030	3.51	0.051	4.39
		0.7	0.039	4.09	0.052	4.40
	(C.2)	0.3	0.037	3.49	0.035	9.03
		0.5	0.042	3.57	0.042	8.91
		0.7	0.041	4.07	0.043	8.96
5000	(5.1)	0.3	0.028	3.41	0.053	5.10
		0.5	0.042	3.64	0.055	5.10
		0.7	0.048	4.18	0.053	5.10
	(C.2)	0.3	0.050	3.84	0.045	10.17
		0.5	0.054	4.00	0.049	10.14
		0.7	0.055	4.15	0.054	10.14

Table J: Testing Parametric Form - Empirical size of our adaptive tests $\widehat{\mathrm{T}}_n$ and of $\widehat{\mathrm{IT}}_n$. Nominal level $\alpha = 0.05$. Design from Appendix C.3.

Horowitz [2006] assumes $d_x = d_w$ and hence we cannot compare our adaptive test with his for Design (C.2). Instead we will compare our adaptive test $\widehat{\mathrm{T}}_n$ against an adaptive image-space test (IT), which is our proposed adaptive version of Bierens [1990]'s type test for semi-nonparametric conditional moment restrictions.[10] Specifically, our image-space test (IT) is based on a leave-one-out sieve estimator of the quadratic functional

---

[10] We refer readers to our first submitted version Breunig and Chen [2020] for the theoretical properties of the adaptive image-space test.
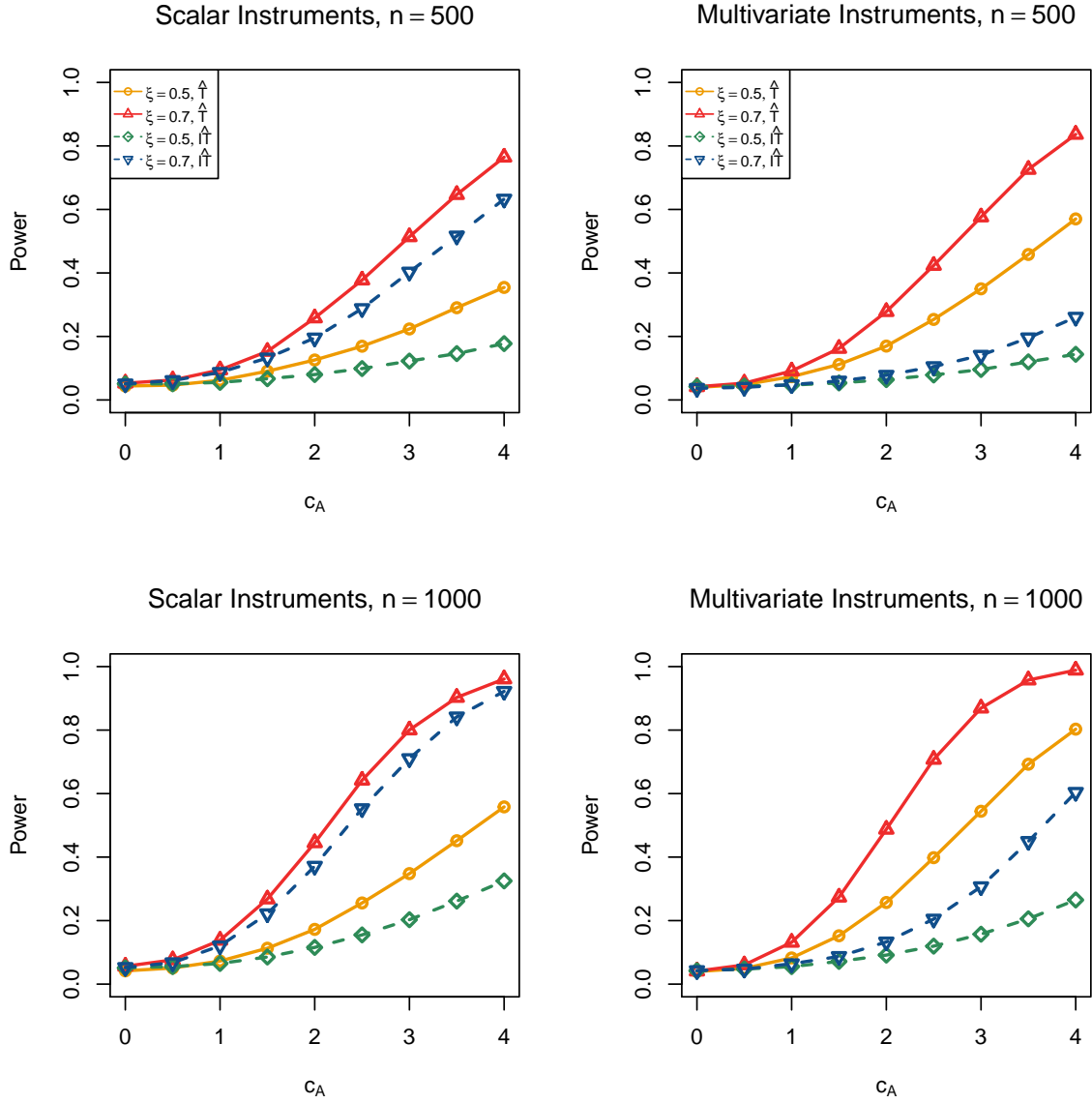
Figure E: Testing Parametric Form - Size-adjusted empirical power of our adaptive tests $\widehat{\mathtt{T}}_n$ (solid lines) and of $\widehat{\mathtt{IT}}_n$ (dashed lines), $\xi = \{0.5, 0.7\}$. LHS: power comparison in scalar case; RHS: power comparison in multivariate case. Design from Appendix C.3.

$\mathrm{E}[\mathrm{E}[Y - h^{\mathrm{R}}(X)|W]^2]$, given by

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

where $\widehat{h}^{\mathrm{R}}$ is a null restricted parametric estimator for the null parametric function h^{R}. The data-driven IT statistic is:

$$\widehat{\mathtt{IT}}_n = \mathbb{1}\left\{\text{there exists } K \in \widehat{\mathcal{I}}_n \text{ such that } \frac{n\widehat{D}_K}{\widehat{v}_K} > \frac{q\big(\alpha/\#(\widehat{\mathcal{I}}_n), K\big) - K}{\sqrt{K}}\right\}$$

with the estimated normalization factor $\widehat{v}_K = \left\| (B'B)^{-1/2}B'\text{DIAG}(Y - \widehat{h}^{\text{R}})^2 B(B'B)^{-1/2} \right\|_F$, and the adjusted index set $\widehat{\mathcal{I}}_n = \left\{ K \leq \widehat{K}_{\max} : \quad K = \underline{K}2^k \text{ where } k = 0, 1, \ldots, k_{\max} \right\}$, where $\underline{K} := \lfloor \sqrt{\log \log n} \rfloor$, $k_{\max} := \lceil \log_2(n^{1/3}/\underline{K}) \rceil$, and the empirical upper bound $\widehat{K}_{\max} = \min \left\{ K > \underline{K} : \quad 1.5\,\zeta^2(K)\sqrt{(\log K)/n} \geq s_{\min}((B'B/n)^{-1/2}) \right\}$. Finally $q(a, K)$ is the $100(1 - a)\%$-quantile of the chi-square distribution with K degrees of freedom.

Table J compares the empirical size of the adaptive image space test $\widehat{\text{IT}}_n$ with our adaptive structural space test $\widehat{\text{T}}_n$, at the 5% nominal level. We see that both tests provide accurate size control. We also report the average choices of sieve dimension parameters, as described in Section 5. The multivariate design (C.2) leads to larger sieve dimension choices $\widehat{K}$ in adaptive image-space tests $\widehat{\text{IT}}_n$ while the sieve dimension choices $\widehat{J}$ of our adaptive structural-space test $\widehat{\text{T}}_n$ is not sensitive to the dimensionality (d_w) of the conditional instruments. Figure E compares the size-adjusted empirical power of $\widehat{\text{IT}}_n$ and of $\widehat{\text{T}}_n$, at the 5% nominal level, using the sample sizes $n = 500$ (first row) and $n = 1000$ (second row). For the scalar conditional instrument case (see the LHS of Figure E), while our adaptive structural space test $\widehat{\text{T}}_n$ is more powerful when $\xi = 0.5$ (weaker strength of instruments), the finite sample power of both tests is similar when $\xi = 0.7$. For the multivariate conditional instruments case (see the RHS of Figure E), while the power of our adaptive structural space test $\widehat{\text{T}}_n$ increases with larger dimension d_w, the adaptive image space test $\widehat{\text{IT}}_n$ suffers from larger d_w and has lower power.

D. Proofs of Inference Results in Subsection 4.3

Proof of Corollary 4.2. Proof of (4.10). We observe

$$\limsup_{n \to \infty} \sup_{h \in \mathcal{H}_0} \text{P}_h\left(h \notin \mathcal{C}_n(\alpha)\right) = \limsup_{n \to \infty} \sup_{h \in \mathcal{H}_0} \text{P}_h\left(\max_{J \in \widehat{\mathcal{I}}_n} \frac{n\widehat{D}_J(h)}{\widehat{\eta}_J(\alpha)\,\widehat{v}_J} > 1\right) \leq \alpha,$$

where the last inequality is due to step 1 of the proof of Theorem 4.1 and step 3 of the proof of Theorem 4.2.

Proof of (4.11). Let J^* be as be as in step 2 of the proof of Theorem 4.1. We observe uniformly for $h \in \mathcal{H}_1(\delta^\circ, \mathsf{r}_n)$ that

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

for n sufficiently large, where the last inequality is due to step 2 of the proof of Theorem 4.1 and step 3 of the proof of Theorem 4.2. $\qquad \square$

Proof of Corollary 4.3. For any $h \in \mathcal{H}_0$, we analyze the diameter of the confidence set $\mathcal{C}_n(\alpha)$ under P_h. Lemma E.10 implies $\sup_{h \in \mathcal{H}_0} \mathrm{P}_h(\widehat{J}_{\max} > \overline{J}) = o(1)$ and hence, it is sufficient to consider the deterministic index set $\mathcal{I}_n$ given in (4.2). For all $h_1 \in \mathcal{C}_n(\alpha) \subset \mathcal{H}_0$ it holds for all $J \in \mathcal{I}_n$ by using the definition of the projection Q_J given in (B.1):

$$
\begin{aligned}
\|h - h_1\|_\mu &\leq \|Q_J \Pi_J (h - h_1)\|_\mu + \|\Pi_J h - h\|_\mu + \|\Pi_J h_1 - h_1\|_\mu \\
&\leq \|Q_J(h - h_1)\|_\mu + O(J^{-p/d_x}),
\end{aligned}
\tag{D.1}
$$

where the second inequality due to the triangular inequality and the sieve approximation bound in Assumption 2(v). The upper bound established in (E.4) yields:

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

with probability approaching one uniformly for $h \in \mathcal{H}_0$. Consequently, the definition of the confidence set $\mathcal{C}_n(\alpha)$ with $h_1 \in \mathcal{C}_n(\alpha)$ gives for all $J \in \mathcal{I}_n$:

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

with probability approaching one uniformly for $h \in \mathcal{H}_0$ by using Lemma E.9 and Lemma E.6(ii). Following the derivation of the upper bound (B.4), there exists a constant $C_B > 0$ such that

$$
n^{-1/2} \big\| \langle Q_J(h - h_1), \psi^J \rangle'_\mu (G_b^{-1/2} S)_l^- \big\| = C_B \Big(n^{-1/2} s_J^{-1} \big(\|h - h_1\|_\mu + J^{-p/d_x} \big) \Big).
$$

Consequently, inequality (D.1) yields

$$
\|h - h_1\|_\mu^2 \leq C \frac{n^{-1} \sqrt{\log \log n} \, \mathrm{v}_J + J^{-2p/d_x}}{1 - C_B n^{-1/2} s_J^{-1}}.
$$

with probability approaching one uniformly for $h \in \mathcal{H}_0$. Now using that $n^{-1/2} s_J^{-1} = o(1)$ for all $J \in \mathcal{I}_n$ by Assumption 4(i) we obtain

$$
\|h - h_1\|_\mu = O \Big(n^{-1/2} (\log \log n)^{1/4} \sqrt{\mathrm{v}_J} + J^{-p/d_x} \Big)
$$

with probability approaching one uniformly for $h \in \mathcal{H}_0$. We may choose $J = J_0 \in \mathcal{I}_n$ for n sufficiently large and hence, the result follows. $\qquad \square$

8

E. Technical Results

Below, $\lambda_{\max}(\cdot)$ denotes the maximal eigenvalue of a matrix.

Theorem E.1. *Let Assumptions 1(i)-(iii) and 2 be satisfied. Then, it holds*

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

Proof. We make use of the decomposition

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

Note that

$$\|Q_J(h - h_0)\|_\mu^2 = \mathrm{E}\left[\left(\psi^J(X)'(G_b^{-1/2}S)_l^-\,\mathrm{E}_h[(Y - h_0(X))\widetilde{b}^K(W)]\right)^2 \mu(X)\right]$$
$$= \left\|G^{1/2}(G_b^{-1/2}S)_l^-\,\mathrm{E}_h[(Y - h_0(X))\widetilde{b}^K(W)]\right\|^2 = \|\,\mathrm{E}_h[V^J]\|^2$$

using the notation $V_i^J = (Y_i - h_0(X_i))G^{1/2}\widetilde{A}\widetilde{b}^K(W_i)$. Thus, the definition of the estimator $\widehat{D}_J$ implies

$$\widehat{D}_J(h_0) - \|Q_J(h - h_0)\|_\mu^2 = \frac{1}{n(n-1)}\sum_{j=1}^J\sum_{i\neq i'}\left(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\right) \qquad (\text{E.1})$$

$$+ \frac{1}{n(n-1)}\sum_{i\neq i'}Y_iY_{i'}b^K(W_i)'\left(A'A - \widehat{A}'\widehat{A}\right)b^K(W_{i'}), \quad (\text{E.2})$$

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

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

We distinguish three different cases. First: i, i', i'', i''' are all different, second: either $i = i''$

9

or $i' = i'''$, or third: $i = i'$ and $i' = i'''$. We thus calculate for each $j, j' \geq 1$ that

$$
\sum_{i \neq i'} \sum_{i'' \neq i'''} \left(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\right)\left(V_{i''j'}V_{i'''j'} - \mathrm{E}_h[V_{1j'}]^2\right)
$$

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

$$
+ 2 \sum_{i \neq i' \neq i''} \left(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\right)\left(V_{i''j'}V_{i'j'} - \mathrm{E}_h[V_{1j'}]^2\right)
$$

$$
+ \sum_{i \neq i'} \left(V_{ij}V_{i'j} - \mathrm{E}_h[V_{1j}]^2\right)\left(V_{ij'}V_{i'j'} - \mathrm{E}_h[V_{1j'}]^2\right).
$$

Due to independent observations we have

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

Consequently, we calculate

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

$$
= 2n(n-1)(n-2)\underbrace{\sum_{j,j'=1}^{J}\mathrm{E}_h\left[\left(V_{1j}V_{2j} - \mathrm{E}_h[V_{1j}]^2\right)\left(V_{3j'}V_{2j'} - \mathrm{E}_h[V_{1j'}]^2\right)\right]}_{I}
$$

$$
+ n(n-1)\underbrace{\sum_{j,j'=1}^{J}\mathrm{E}_h\left[\left(V_{1j}V_{2j} - \mathrm{E}_h[V_{1j}]^2\right)\left(V_{1j'}V_{2j'} - \mathrm{E}_h[V_{1j'}]^2\right)\right]}_{II}.
$$

To bound the summand I we observe that

$$
I = \sum_{j,j'=1}^{J}\mathrm{E}_h[V_{1j}]\,\mathrm{E}_h[V_{1j'}]\,\mathbb{C}ov_h(V_{1j}, V_{1j'}) = \mathrm{E}_h[V_1^J]'\,\mathbb{C}ov_h(V_1^J, V_1^J)\,\mathrm{E}_h[V_1^J]
$$

$$
\leq \lambda_{\max}\left(\mathbb{V}ar_h((Y - h_0(X))\widetilde{b}^K(W))\right)\left\|(G_b^{-1/2}SG^{-1/2})_l^{-}\,\mathrm{E}_h[V_1^J]\right\|^2
$$

$$
\leq \bar{\sigma}^2\left\|\left((G_b^{-1/2}S)_l^{-}\,\mathrm{E}_h[(Y - h_0(X))\widetilde{b}^K(W)]\right)' G(G_b^{-1/2}S)_l^{-}\right\|^2
$$

$$
= \bar{\sigma}^2\left\|\langle Q_J(h - h_0), \psi^J\rangle'_{\mu}(G_b^{-1/2}S)_l^{-}\right\|^2
$$

by using the notation $V_i^J = (Y_i - h_0(X_i))(G_b^{-1/2}SG^{-1/2})_l^{-}\widetilde{b}^K(W_i)$ and Lemma E.7, 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

$$II = n(n-1) \sum_{j,j'=1}^{J} \mathrm{E}_h[V_{1j}V_{1j'}]^2 - n(n-1)\Big(\sum_{j=1}^{J} \mathrm{E}_h[V_{1j}]^2 \Big)^2$$

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

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

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

$$\leq 2\overline{\sigma}^2 \left(\frac{1}{n} \big\| \langle Q_J(h-h_0), \psi^J \rangle'_\mu (G_b^{-1/2}S)_l^- \big\|^2 + \frac{\mathrm{v}_J^2}{n^2} \right). \tag{E.3}$$

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

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

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

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

which completes the proof. $\qquad\square$

Lemma E.1. *Let Assumptions 2(i) be satisfied. Then, it holds uniformly for $h \in \mathcal{H}$ and uniformly for $J \in \mathcal{I}_n$:*

$$\mathrm{v}_J \leq \overline{\sigma}^2 s_J^{-2} \sqrt{J}.$$

Proof. Note that for any $J \times J$ matrix m it holds $\|m\|_F \leq \sqrt{J}\|m\|$ and hence

$$\mathrm{v}_J^2 = \left\| \big(G_b^{-1/2}SG^{-1/2}\big)_l^- \mathrm{E}\left[(Y - h(X))^2 \widetilde{b}^K(W)\widetilde{b}^K(W)'\right] \big(G_b^{-1/2}SG^{-1/2}\big)_l^- \right\|_F^2$$

$$\leq J \left\| \big(G_b^{-1/2}SG^{-1/2}\big)_l^- \right\|^4 \left\| \mathrm{E}\left[(Y - h(X))^2 \widetilde{b}^K(W)\widetilde{b}^K(W)'\right] \right\|^2.$$

Consequently, the result follows from using the relationship $\|(G_b^{-1/2}SG^{-1/2})_l^-\| = s_J^{-1}$ and Lemma E.7, i.e., $\|\mathrm{E}[(Y - h(X))^2\widetilde{b}^K(W)\widetilde{b}^K(W)']\| \leq \overline{\sigma}^2$ uniformly for $h \in \mathcal{H}$ and uniformly for $J \in \mathcal{I}_n$. $\qquad\square$

Lemma E.2. *Let Assumptions 1(i) and 4(i) be satisfied. Then, it holds uniformly for $h \in \mathcal{H}$ and uniformly for $J \in \mathcal{I}_n$:*

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

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

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

$$Q' A G_b A' Q = \mathrm{diag}(s_1^{-2}, \ldots, s_J^{-2}).$$

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

$$v_J^2 = \sum_{j,j'=1}^{J} \mathrm{E}[\widetilde{V}_{1j} \widetilde{V}_{1j'}]^2 \geq \sum_{j=1}^{J} \mathrm{E}[\widetilde{V}_{1j}^2]^2 = \sum_{j=1}^{J} \left(\mathrm{E} \, |(Y - h(X)) e_j' Q' A b^K(W)|^2 \right)^2.$$

Consequently, using the lower bound $\inf_{w \in \mathcal{W}} \inf_{h \in \mathcal{H}} \mathrm{E}[(Y - h(X))^2 | W = w] \geq \underline{\sigma}^2$ by Assumption 1(i), we obtain uniformly for $h \in \mathcal{H}$:

$$v_J^2 \geq \underline{\sigma}^4 \sum_{j=1}^{J} \left(\mathrm{E}[e_j' Q' A b^K(W) b^K(W)' A' Q e_j] \right)^2 = \underline{\sigma}^4 \sum_{j=1}^{J} \left(e_j' Q' A G_b A' Q e_j \right)^2$$

$$= \underline{\sigma}^4 \sum_{j=1}^{J} \left(e_j' \, \mathrm{diag}(s_1^{-2}, \ldots, s_J^{-2}) e_j \right)^2 \geq \underline{\sigma}^4 \sum_{j=1}^{J} s_j^{-4},$$

which proves the result. $\qquad\square$

Lemma E.3. *Let Assumption 2 be satisfied. Then we have uniformly in $h \in \mathcal{H}$,*

$$\|Q_J(h - \Pi_{\mathrm{R}} h)\|_\mu = \|h - \Pi_{\mathrm{R}} h\|_\mu + O\big(J^{-p/d_x}\big).$$

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

$$\|Q_J(h - \Pi_{\mathrm{R}} h)\|_\mu = \left\| (G_b^{-1/2} S G^{-1/2})_l^- \, \mathrm{E}[\widetilde{b}^K(W)(h - \Pi_{\mathrm{R}} h)(X)] \right\|$$

$$\leq \left\| (G_b^{-1/2} S G^{-1/2})_l^- \, \mathrm{E}[\widetilde{b}^K(W)(\Pi_J h - \Pi_J \Pi_{\mathrm{R}} h)(X)] \right\|$$

$$+ \left\| (G_b^{-1/2} S G^{-1/2})_l^- \, \mathrm{E}[\widetilde{b}^K(W)((h - \Pi_{\mathrm{R}} h)(X) - (\Pi_J h - \Pi_J h_0)(X))] \right\|$$

$$\leq \|\Pi_J h - \Pi_J \Pi_{\mathrm{R}} h\|_\mu + s_J^{-1} \|\Pi_K T((h - \Pi_{\mathrm{R}} h) - (\Pi_J h - \Pi_J \Pi_{\mathrm{R}} h))\|_{L^2(W)}$$

$$\leq \|\Pi_J h - \Pi_J \Pi_{\mathrm{R}} h\|_\mu + O\big(J^{-p/d_x}\big)$$

by making use of Assumption 2(iv) and the sieve approximation bound in Assumption 2(v). $\qquad \square$

Lemma E.4. *Let Assumptions 1(i)-(iii) and 2 hold. Then, uniformly in $h \in \mathcal{H}$ it holds*

$$\frac{1}{n(n-1)} \sum_{i \neq i'} \left(Y_i - h_0(X_i)\right)\left(Y_{i'} - h_0(X_{i'})\right) b^K(W_i)' \left(A'A - \widehat{A}'\widehat{A}\right) b^K(W_{i'})$$
$$= O_p\left(n^{-1} v_J + n^{-1/2} \|\langle Q_J(h - h_0), \psi^J \rangle'_\mu (G_b^{-1/2} S)_l^- \|\right).$$

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

$$\frac{1}{n^2} \sum_{i,i'} \left(Y_i - h_0(X_i)\right)\left(Y_{i'} - h_0(X_{i'})\right) b^K(W_i)' \left(A'A - \widehat{A}'\widehat{A}\right) b^K(W_{i'})$$

$$= \mathrm{E}[(h - h_0)(X) b^K(W)]' \left(A'A - \widehat{A}'\widehat{A}\right) \mathrm{E}[(h - h_0)(X) b^K(W)]$$

$$+ 2\left(\frac{1}{n} \sum_i \left(Y_i - h_0(X_i)\right) b^K(W_i)' - \mathrm{E}[(h - h_0)(X) b^K(W)]'\right) \left(A'A - \widehat{A}'\widehat{A}\right)$$

$$\times \mathrm{E}[(h - h_0)(X) b^K(W)]$$

$$+ \left(\frac{1}{n} \sum_i \left(Y_i - h_0(X_i)\right) b^K(W_i)' - \mathrm{E}[(h - h_0)(X) b^K(W)]'\right) \left(A'A - \widehat{A}'\widehat{A}\right)$$

$$\times \left(\frac{1}{n} \sum_i \left(Y_i - h_0(X_i)\right) b^K(W_i)' - \mathrm{E}[(h - h_0)(X) b^K(W)]'\right)$$

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

$$\mathrm{E}[(h - h_0)(X) b^K(W)]' \left(A'A - \widehat{A}'\widehat{A}\right) \mathrm{E}[(h - h_0)(X) b^K(W)]$$

$$= 2 \mathrm{E}[(h - h_0)(X) b^K(W)]' A'(A - \widehat{A}) \mathrm{E}[(h - h_0)(X) b^K(W)]$$

$$- \mathrm{E}[(h - h_0)(X) b^K(W)]' (A - \widehat{A})'(A - \widehat{A}) \mathrm{E}[(h - h_0)(X) b^K(W)]$$

$$= 2T_1 - T_2,$$

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

$$T_1 = O_p\left(n^{-1/2} \|\langle Q_J(h - h_0), \psi^J \rangle'_\mu (G_b^{-1/2} S)_l^- \|\right). \tag{E.5}$$

To do so, we make use of the decomposition

$$T_1 = \mathrm{E}[(h - h_0)(X) b^K(W)]' A'(\widehat{A} - A) \mathrm{E}[\Pi_J(h - h_0)(X) b^K(W)]$$
$$+ \mathrm{E}[(h - h_0)(X) b^K(W)]' A'(\widehat{A} - A) \mathrm{E}[(h - h_0 - \Pi_J(h - h_0))(X) b^K(W)]. \tag{E.6}$$

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

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

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

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

where the second bound is due to the Cauchy-Schwarz inequality and the third bound is due to Assumption 2(iv). To establish an upper bound for T_{12} we infer from Chen and Christensen [2018, Lemma F.10(c)] that

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

using Assumption 2(ii), i.e., $s_J^{-1}\zeta_J^2\sqrt{(\log J)/n} = O(1)$. Consider the second summand on

the right hand side of (E.6). Following the upper bound of T_{12} we obtain

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

using that $s_J^{-2}\|T(h - h_0 - \Pi_J(h - h_0))\|^2_{L^2(W)} = O(\|h - h_0 - \Pi_J(h - h_0)\|^2_\mu)$ by Assumption 2(iv) and $\zeta_J^2(\log J)\|h - \Pi_J h\|^2_\mu = O(1)$ by Assumption 2(iii) and the sieve approximation bound in Assumption 2(v), which implies the upper bound (E.5).

Consider T_2. We make use of the decomposition

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

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

$$\mathrm{E}\left\| \left(\frac{1}{n}\sum_i (h - h_0)(X_i)b^K(W_i) - \mathrm{E}[(h - h_0)(X)b^K(W)] \right)' A'G^{1/2} \right\|^2$$
$$\leq n^{-1}\,\mathrm{E}\left[(h - h_0(X))^2 \|b^K(W)'A'G^{1/2}\|^2 \right] \leq n^{-1}s_J^{-2}\sqrt{J},$$

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

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

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. $\square$

Lemma E.5. *Let Assumptions 1(i)-(iii), 2(i)(v) and 4(i) hold. Then, under $\mathcal{H}_0 = \{h_0\}$ it holds uniformly in $J \in \mathcal{I}_n$:*

$$\frac{1}{n-1}\sum_{i \neq i'}\big(Y_i - h_0(X_i)\big)\big(Y_{i'} - h_0(X_{i'})\big)b^K(W_i)'\Big(A'A - \widehat{A}'\widehat{A}\Big)b^K(W_{i'}) = o_p\left(v_J\sqrt{\log\log J}\right).$$

Proof. Let I_{s_J} denote the $K \times K$ dimensional identity matrix multiplied by the vector $c_0(s_1,\ldots,s_J)$ for some sufficiently large constant c_0 and where s_j^{-1}, $1 \leq j \leq J$, are the nondecreasing singular values of $AG_b^{1/2} = (G_b^{-1/2}SG^{-1/2})_l^-$. We make use of the inequality

$$\sum_{i \neq i'}\big(Y_i - h_0(X_i)\big)\big(Y_{i'} - h_0(X_{i'})\big)b^K(W_i)'\Big(A'A - \widehat{A}'\widehat{A}\Big)b^K(W_{i'})$$

$$\leq \left\|\sum_i\big(Y_i - h_0(X_i)\big)\widetilde{b}^K(W_i)I_{s_J}^{-1}\right\|^2\left\|I_{s_J}G_b^{1/2}(A'A - \widehat{A}'\widehat{A})G_b^{1/2}I_{s_J}\right\|$$

$$= \sum_{i \neq i'}\big(Y_i - h_0(X_i)\big)\big(Y_{i'} - h_0(X_{i'})\big)\widetilde{b}^K(W_i)'I_{s_J}^{-2}\widetilde{b}^K(W_{i'})\left\|I_{s_J}G_b^{1/2}(A'A - \widehat{A}'\widehat{A})G_b^{1/2}I_{s_J}\right\|$$

$$+ \sum_i\left\|\big(Y_i - h_0(X_i)\big)\widetilde{b}^K(W_i)I_{s_J}^{-1}\right\|^2\left\|I_{s_J}G_b^{1/2}(A'A - \widehat{A}'\widehat{A})G_b^{1/2}I_{s_J}\right\|.$$

The fourth moment condition imposed in Assumption 2(i) implies

$$\mathrm{E}\max_{J \in \mathcal{I}_n}\left|\frac{1}{nv_J}\sum_i\left(\|(Y_i - h_0(X_i))\widetilde{b}^K(W_i)I_{s_J}^{-1}\|^2 - \mathrm{E}\,\|(Y - h_0(X))\widetilde{b}^K(W)I_{s_J}^{-1}\|^2\right)\right|^2$$

$$\leq Cn^{-1}\sum_{J \in \mathcal{I}_n}v_J^{-1}\,\mathrm{E}\,\|\widetilde{b}^K(W)I_{s_J}^{-1}\|^4$$

$$\leq Cn^{-1}\zeta_{\overline{J}}\sum_{J \in \mathcal{I}_n}v_J^{-1}\,\mathrm{E}\,\|\widetilde{b}^K(W)I_{s_J}^{-2}\|^2$$

$$\leq Cc_0^{-1}n^{-1}\zeta_{\overline{J}}\sum_{J \in \mathcal{I}_n}v_J^{-1}\sum_{j=1}^J s_j^{-4}$$

$$\leq Cc_0^{-1}\underline{\sigma}^{-4}n^{-1}\zeta_{\overline{J}}\#(\mathcal{I}_n) = o(1),$$

where the last inequality is due to Lemma E.2 and the definition of the index set $\mathcal{I}_n$ which implies $n^{-1/2}\#(\mathcal{I}_n) = o(1)$. Consequently, due to the second moment condition imposed in Assumption 2(i) we obtain uniformly for $J \in \mathcal{I}_n$

$$n^{-1}\sum_i\|(Y_i - h_0(X_i))\widetilde{b}^K(W_i)I_{s_J}^{-1}\|^2 \leq \overline{\sigma}^2c_0^{-1}\zeta_J\sqrt{\sum_{j=1}^J s_j^{-4}} \leq \overline{\sigma}^2\underline{\sigma}^{-2}c_0^{-1}\zeta_Jv_J$$

with probability approaching one, by making use of Lemma E.2. Further, we obtain

$$
P_{h_0}\left(\max_{J\in\mathcal{I}_n}\left|\frac{(\log\log J)^{-1/2}}{(n-1)v_J}\sum_{i,i'}\left(Y_i-h_0(X_i)\right)\left(Y_{i'}-h_0(X_{i'})\right)b^K(W_i)'\left(A'A-\widehat{A}'\widehat{A}\right)b^K(W_{i'})\right|>1\right)
$$

$$
\leq P_{h_0}\left(\max_{J\in\mathcal{I}_n}\left|\frac{(\log\log J)^{-1/2}}{(n-1)v_J}\sum_{i\neq i'}\left(Y_i-h_0(X_i)\right)\left(Y_{i'}-h_0(X_{i'})\right)\widetilde{b}^K(W_i)'I_{s_J}^{-2}\widetilde{b}^K(W_{i'})\right|>1\right)
$$

$$
+P_{h_0}\left(\max_{J\in\mathcal{I}_n}\left(\|I_{s_J}G_b^{1/2}(A'A-\widehat{A}'\widehat{A})G_b^{1/2}I_{s_J}\|\right)>1/2\right)
$$

$$
+P_{h_0}\left(\max_{J\in\mathcal{I}_n}\left(\overline{\sigma}^2\underline{\sigma}^{-2}c_0^{-1}\zeta_J(\log\log J)^{-1/2}\|I_{s_J}G_b^{1/2}(A'A-\widehat{A}'\widehat{A})G_b^{1/2}I_{s_J}\|\right)>1/2\right)+o(1)
$$

$$
=T_1+T_2+T_3+o(1).
$$

Note that T_1 is arbitrarily small for c_0 sufficiently large by following step 1 in the proof of Theorem 4.1. Consider T_2. We make use of the inequality

$$
\left\|I_{s_J}G_b^{1/2}(\widehat{A}'\widehat{A}-A'A)G_b^{1/2}I_{s_J}\right\|\leq 2\left\|I_{s_J}G_b^{1/2}(\widehat{A}-A)'AG_b^{1/2}I_{s_J}\right\|+\left\|(\widehat{A}-A)G_b^{1/2}I_{s_J}\right\|^2.
$$

It is sufficient to consider the first summand on the right hand side. Note that $\|AG_b^{1/2}I_{s_J}\|\leq c_0^{-1}$. Consequently, we obtain uniformly for $J\in\mathcal{I}_n$ that by making use of the upper bound (B.5):

$$
\left\|I_{s_J}G_b^{1/2}(\widehat{A}-A)'AG_b^{1/2}I_{s_J}\right\|\leq c_0^{-1}s_J\left\|(\widehat{G}_b^{-1/2}\widehat{S}\widehat{G}^{-1/2})_l^-\widehat{G}_b^{-1/2}G_b^{1/2}-(G_b^{-1/2}SG^{-1/2})_l^-\right\|
$$

$$
=O_p\left(s_J^{-1}\zeta_J\sqrt{(\log J)/n}\right)=o_p(1),
$$

where the last equation is due to Assumption 4(i), i.e., $s_J^{-1}\zeta_J^2\sqrt{(\log J)/n}=O(1)$ uniformly for $J\in\mathcal{I}_n$. This rate condition immediately implies $T_3=o(1)$. □

Lemma E.6. *Let Assumption 1(i)-(iii) be satisfied.*
(i) If in addition Assumption 2 holds, then for any $c>0$ we have

$$
\sup_{h\in\mathcal{H}}P_h\left(\left|1-\frac{\widehat{v}_J}{v_J}\right|>c\right)=o(1).
$$

(ii) If in addition Assumptions 2(i)(v) and 4(i) hold, then for any $c>0$ we have

$$
\sup_{h\in\mathcal{H}}P_h\left(\max_{J\in\mathcal{I}_n}\left|1-\frac{\widehat{v}_J}{v_J}\right|>c\right)=o(1).
$$

Proof. It is sufficient to prove (ii). We denote $\Sigma=\mathrm{E}[(Y-h(X))^2b^K(W)b^K(W)']$ and its empirical analog $\widehat{\Sigma}=n^{-1}\sum_i\left(Y_i-\widehat{h}_J(X_i)\right)^2b^K(W_i)b^K(W_i)'$. Hence, for all $J\in\mathcal{I}_n$ the

triangular inequality implies

$$|\widehat{v}_J - v_J| \le \left\|\widehat{A}\widehat{\Sigma}\widehat{A}' - A\Sigma A'\right\|_F$$

$$\le 2\left\|(\widehat{A} - A)\widehat{\Sigma}\,A'\right\|_F + \left\|(\widehat{A} - A)\widehat{\Sigma}^{1/2}\right\|_F^2 + \left\|A(\widehat{\Sigma} - \Sigma)A'\right\|_F.$$

In the remainder of this proof, it is sufficient to consider

$$\left\|(\widehat{A} - A)\Sigma\,A'\right\|_F + \left\|A(\widehat{\Sigma} - \Sigma)\,A'\right\|_F = T_1 + T_2.$$

Consider T_1. By Lemma E.7 we have the upper bound $\|G_b^{-1/2}\Sigma G_b^{-1/2}\| \le \overline{\sigma}$. Below we make use of the inequality $\|m_1 m_2\|_F \le \|m_1\|\|m_2\|_F$ for matrices m_1 and m_2. Since the Frobenius norm is invariant under rotation, we calculate uniformly for $J \in \mathcal{I}_n$ that

$$T_1 = \|(G_b^{1/2}SG^{1/2})(\widehat{A} - A)\Sigma\,A'AG_b^{1/2}\|$$

$$\le \|(G_b^{1/2}SG^{1/2})(\widehat{A} - A)G_b^{1/2}\|\|G_b^{-1/2}\Sigma G_b^{-1/2}\|\|(G_b^{1/2}SG^{1/2})_l^{-2}\|_F$$

$$= O_p\Big(s_J^{-1}\zeta_J\Big(n^{-1}(\log J)\sum_{j=1}^{J}s_j^{-4}\Big)^{1/2}\Big)$$

by making use of the upper bound (B.5) and the Schur decomposition as in the proof of Lemma E.2. From Assumption 4(i), i.e., $s_J^{-1}\zeta_J^2\sqrt{(\log J)/n} = O(1)$ uniformly for $J \in \mathcal{I}_n$, we infer

$$T_1 = O_p\Big(J^{-1/2}\Big(\sum_{j=1}^{J}s_j^{-4}\Big)^{1/2}\Big) = o_p(v_J)$$

uniformly for $J \in \mathcal{I}_n$, where the last equation is due to Lemma E.2. Consider T_2. Again using Lemma E.2 we obtain

$$T_2 \le \underline{\sigma}^{-2}\|G_b^{-1/2}(\widehat{\Sigma} - \Sigma)G_b^{-1/2}\|$$

by using the upper bound as derived for T_1. Further, evaluate

$$\|G_b^{-1/2}(\widehat{\Sigma} - \Sigma)G_b^{-1/2}\| = \left\|\frac{1}{n}\sum_i \big((Y_i - \widehat{h}_J(X_i))^2 - (Y_i - h(X_i))^2\big)\widetilde{b}^K(W_i)\widetilde{b}^K(W_i)'\right\|$$

$$\le \left\|\frac{1}{n}\sum_i \big(\widehat{h}_J(X_i) - h(X_i)\big)^2\widetilde{b}^K(W_i)\widetilde{b}^K(W_i)'\right\|$$

$$+ 2\left\|\frac{1}{n}\sum_i \big(\widehat{h}_J(X_i) - h(X_i)\big)(Y_i - h(X_i))\widetilde{b}^K(W_i)\widetilde{b}^K(W_i)'\right\|$$

$$= T_{21} + T_{22}.$$

Consider T_{21}. The definition of the unrestricted sieve NPIV estimator in (2.3) implies uniformly for $J \in \mathcal{I}_n$

$$
\begin{aligned}
T_{21} &\leq \left\| \frac{1}{n} \sum_i \big(\widehat{h}_J(X_i) - Q_J h(X_i) \big)^2 \widetilde{b}^K(W_i) \widetilde{b}^K(W_i)' \right\| \\
&\quad + \left\| \frac{1}{n} \sum_i \big(Q_J h(X_i) - h(X_i) \big)^2 \widetilde{b}^K(W_i) \widetilde{b}^K(W_i)' \right\| \\
&\leq \zeta_J^2 \left\| \widehat{A} \frac{1}{n} \sum_i Y_i b^K(W_i) - A \operatorname{E}[Y b^K(W)] \right\|^2 \times \left\| \frac{1}{n} \sum_i \psi^J(X_i) \psi^J(X_i)' \right\| \\
&\quad + \zeta_J^2 \left\| \frac{1}{n} \sum_i \big(Q_J h(X_i) - h(X_i) \big)^2 \right\| \\
&= O_p \left(\zeta_J^4 s_J^{-2} n^{-1} + \max_{J \in \mathcal{I}_n} \big\{ \zeta_J^2 \| Q_J h - h \|_{L^2(X)} \big\} \right) = o_p(1)
\end{aligned}
$$

uniformly in $h \in \mathcal{H}$. The last equation follows by the rate condition imposed in Assumption 4(i) and that $\| Q_J h - h \|_{L^2(X)} \leq \| Q_J h - h \|_\mu = O(J^{-p/d_x})$ uniformly for $J \in \mathcal{I}_n$ and $h \in \mathcal{H}$ by following the proof of Lemma E.3. Analogously, we obtain $T_{22} = o_p(1)$ uniformly for $J \in \mathcal{I}_n$. $\qquad \square$

Lemma E.7. *Under Assumptions 2(i) it holds*

$$
\sup_{J \in \mathcal{I}_n} \sup_{h \in \mathcal{H}} \lambda_{\max} \big(\operatorname{E}_h \big[(Y - \Pi_{\mathrm{R}} h(X))^2 \widetilde{b}^{K(J)}(W) \widetilde{b}^{K(J)}(W)' \big] \big) \leq \overline{\sigma}^2 < \infty.
$$

Proof. We have for any $\gamma \in \mathbb{R}^K$ where $K = K(J)$ that

$$
\begin{aligned}
\gamma' \operatorname{E}_h \big[(Y - \Pi_{\mathrm{R}} h(X))^2 \widetilde{b}^K(W) \widetilde{b}^K(W)' \big] \gamma &\leq \operatorname{E} \Big[\operatorname{E}_h[(Y - \Pi_{\mathrm{R}} h(X))^2 | W] \big(\gamma' \widetilde{b}^K(W) \big)^2 \Big] \\
&\leq \overline{\sigma}^2 \operatorname{E} \Big[\big(\gamma' \widetilde{b}^K(W) \big)^2 \Big] \\
&= \overline{\sigma}^2 \gamma' G_b^{-1/2} \operatorname{E} \big[b^K(W) b^K(W)' \big] G_b^{-1/2} \gamma = \overline{\sigma}^2 \| \gamma \|^2
\end{aligned}
$$

uniformly for $h \in \mathcal{H}$ and uniformly for $J \in \mathcal{I}_n$, where the second inequality is due to Assumption 2(i). $\qquad \square$

Lemma E.8. *Let Assumptions 1(i)-(iii) and 2 be satisfied. Then, under the simple hypothesis $\mathcal{H}_0 = \{h_0\}$ for a known function h_0, we have*

$$
\limsup_{n \to \infty} \operatorname{P}_{h_0} \left(\frac{n \widehat{D}_J(h_0)}{\widehat{v}_J} > \eta_J(\alpha) \right) \leq \alpha
$$

Proof. Making use of decomposition (E.1–E.2) together with Lemma E.4 yields

$$P_{h_0}\left(\frac{n\widehat{D}_J(h_0)}{v_J} > \eta_J(\alpha)\right) = P_{h_0}\left(\frac{1}{v_J(n-1)}\sum_{j=1}^{J}\sum_{i\neq i'}V_{ij}V_{i'j} > \eta_J(\alpha)\right) + o(1).$$

Using the martingale central limit theorem (see for instance Breunig [2020, Lemma A.3] for more details) we obtain

$$P_{h_0}\left(\frac{1}{\sqrt{2}v_J(n-1)}\sum_{j=1}^{J}\sum_{i\neq i'}V_{ij}V_{i'j} > z_{1-\alpha}\right) = \alpha + o(1),$$

where $z_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the standard normal distribution. Further, Lemma E.6 implies $v_J/\widehat{v}_J = 1 + o_p(1)$ and since $\eta_J(\alpha)/\sqrt{2} = \frac{q(\alpha,J)-J}{\sqrt{2J}}$ converges to $z_{1-\alpha}$ as J tends to infinity, the result follows. $\qquad\square$

Lemma E.9. *Let Assumption 5(iii) be satisfied. Then, for all $\alpha \in (0,1)$ and $J \in \mathcal{I}_n$ we have for n sufficiently large:*

$$\inf_{h\in\mathcal{H}} P_h\left(\forall J \in \mathcal{I}_n: \frac{\sqrt{\log\log(\widehat{\gamma}_J) - \log(\alpha)}}{4} \leq \widehat{\eta}_J(\alpha) \leq 4\sqrt{\log\log(n) - \log(\alpha)}\right) = 1 + o(1).$$

Proof. We first prove the lower bound. By Assumption 5(iii) and the definition of the index set $\mathcal{I}_n$ we have that $\widehat{\gamma}_J$, $J \in \mathcal{I}_n$, tends slowly to infinity as $n \to \infty$ with probability approaching one uniformly for $h \in \mathcal{H}$. From the lower bounds for quantiles of the chi-squared distribution established in Inglot [2010, Theorem 5.2] we deduce for $\widehat{\gamma}_J$ sufficiently large

$$\widehat{\eta}_J(\alpha) = \frac{q(\alpha/\#(\mathcal{I}_n), \widehat{\gamma}_J) - \widehat{\gamma}_J}{\sqrt{\widehat{\gamma}_J}} \geq \frac{\sqrt{\log\left(\#(\mathcal{I}_n)/\alpha\right)}}{4} + \frac{2\log\left(\#(\mathcal{I}_n)/\alpha\right)}{\sqrt{\widehat{\gamma}_J}}.$$

There exists some integer $\bar{j}$ such that $\underline{J}2^{\bar{j}} \leq \overline{J}$. Consequently, the definition of the index set implies for all $J \in \mathcal{I}_n$ that

$$\log(\widehat{\gamma}_J) \leq \log(J) \leq \log(\underline{J}2^{\bar{j}}) = \bar{j}\log(2) + \log(\underline{J}) \leq \bar{j} + 1 = \#(\mathcal{I}_n)$$

for n sufficiently large. Consequently, we obtain

$$\widehat{\eta}_J(\alpha) \geq \frac{\sqrt{\log(\log(\widehat{\gamma}_J)/\alpha)}}{4} + \frac{2\log(\widehat{\gamma}_J)}{\sqrt{\widehat{\gamma}_J}} \geq \frac{\sqrt{\log\log(\widehat{\gamma}_J) - \log(\alpha)}}{4}$$

with probability approaching one uniformly for $h \in \mathcal{H}$. We now consider the upper bound.

From Laurent and Massart [2000, Lemma 1] we deduce the upper bound:

$$\widehat{\eta}_J(\alpha) \leq 2\sqrt{\log\left(\#(\mathcal{I}_n)/\alpha\right)} + \frac{2\log\left(\#(\mathcal{I}_n)/\alpha\right)}{\sqrt{\widehat{\gamma}_J}} \leq 2\sqrt{\log\left(\#(\mathcal{I}_n)/\alpha\right)}(1 + o(1))$$

as $\widehat{\gamma}_J$ tends to infinity. From the definition of $\#(\mathcal{I}_n)$ we infer

$$\#(\mathcal{I}_n) = \overline{j} + 1 \leq \lceil \log_2(n^{1/3}/\underline{J}) \rceil + 1 \leq \log(n^{1/3}/\underline{J}) + 1 \leq \log(n)$$

and hence, we conclude

$$\widehat{\eta}_J(\alpha) \leq 2\sqrt{\log(\log(n)/\alpha)}(1 + o(1)) \leq 4\sqrt{\log\log(n) - \log(\alpha)}$$

with probability approaching one uniformly for $h \in \mathcal{H}$. □

Lemma E.10. *Let Assumption 4(i)(iii) be satisfied. Then $\widehat{J}_{\max}$ given in (2.7) satisfies*

(i) $\quad \sup_{h \in \mathcal{H}} \mathrm{P}_h\left(\widehat{J}_{\max} > \overline{J}\right) = o(1)$ *and*

(ii) $\quad \sup_{h \in \mathcal{H}} \mathrm{P}_h\left(2J_0 > \widehat{J}_{\max}\right) = o(1)$.

Proof. Recall the definition of $\overline{J} = \sup\{J : \zeta^2(J)\sqrt{(\log J)/n} \leq \overline{c}\, s_J\}$. Following the proof of Chen et al. [2021, Lemma C.6], using Weyl's inequality (see e.g. Chen and Christensen [2018, Lemma F.1]) together with Chen and Christensen [2018, Lemma F.7] we obtain that $|\widehat{s}_J - s_J| \leq c_0 s_J$ uniformly in $J \in \mathcal{I}_n$ for some $0 < c_0 < 1$ with probability approaching one uniformly for $h \in \mathcal{H}$.

Proof of (i). By making use of the definition of $\widehat{J}_{\max}$ given in (2.7), we obtain uniformly for $h \in \mathcal{H}$:

$$\mathrm{P}_h\left(\widehat{J}_{\max} > \overline{J}\right) \leq \mathrm{P}_h\left(\zeta^2(\overline{J})\sqrt{\log(\overline{J})/n} < \frac{3}{2}\widehat{s}_{\overline{J}}\right)$$

$$\leq \mathrm{P}_h\left(\zeta^2(\overline{J})\sqrt{\log(\overline{J})/n} < \frac{3}{2}(1 + c_0)s_{\overline{J}}\right) + o(1)$$

The upper bound imposed on the growth of $\overline{J}$ is determined by a sufficiently large constant $\overline{c} > 0$ and hence, there exists a constant $\underline{c} \geq 3(1 + c_0)/2$ such that $s_{\overline{J}}^{-1}\zeta^2(\overline{J})\sqrt{\log(\overline{J})/n} \geq \underline{c}$. Consequently, we obtain

$$\mathrm{P}_h\left(\widehat{J}_{\max} > \overline{J}\right) \leq \mathrm{P}_h\left(s_{\overline{J}}^{-1}\zeta^2(\overline{J})\sqrt{\log(\overline{J})/n} < \frac{3}{2}(1 + c_0)\right) + o(1) = o(1).$$

Proof of (ii). From the definition of J_0 given in (4.3) we infer as above for some constant

$0 < c_0 < 1$ and uniformly for $h \in \mathcal{H}$:

$$\mathrm{P}_h \left(J_0 > \widehat{J}_{\max} \right) \leq \mathrm{P}_h \left((1 - c_0) n^{-1} \sqrt{\log \log n} \, \widehat{J}_{\max}^{2p/d_x + 1/2} \leq \widehat{s}_{\widehat{J}_{\max}}^2 \right) + o(1).$$

Consider the case $\zeta(J) = \sqrt{J}$. The definition of $\widehat{J}_{\max}$ in (2.7) yields uniformly for $h \in \mathcal{H}$:

$$\mathrm{P}_h \left(J_0 > \widehat{J}_{\max} \right) \leq \mathrm{P}_h \left((1 - c_0) \sqrt{\log \log n} \, \widehat{J}_{\max}^{2p/d_x - 3/2} \leq (\log \overline{J}) \right) + o(1)$$

$$\leq \mathrm{P}_h \left((1 - c_0) \widehat{s}_{\widehat{J}_{\max}} \sqrt{n} \leq \frac{2}{3} \sqrt{\log \overline{J}} \left(\frac{\log \overline{J}}{\sqrt{\log \log n}} \right)^{1/(2p/d_x - 3/2)} \right) + o(1)$$

$$\leq \mathrm{P}_h \left((1 - c_0)^2 s_{\overline{J}} \sqrt{n} \leq \frac{2}{3} \sqrt{\log \overline{J}} \left(\frac{\log \overline{J}}{\sqrt{\log \log n}} \right)^{1/(2p/d_x - 3/2)} \right) + o(1)$$

$$\leq \mathrm{P}_h \left(\frac{(1 - c_0)^2}{\overline{c}} \overline{J} \leq \frac{2}{3} \left(\frac{\log \overline{J}}{\sqrt{\log \log n}} \right)^{1/(2p/d_x - 3/2)} \right) + o(1),$$

where the last inequality follows from the definition of $\overline{J}$, i.e., $s_{\overline{J}} \geq \overline{c}^{-1} \overline{J} \sqrt{\log(\overline{J})/n}$. From Assumption 4(iii), i.e., $p \geq 3d_x/4$, we infer $\mathrm{P}_h(J_0 > \widehat{J}_{\max}) = o(1)$ and, in particular, $\mathrm{P}_h(2J_0 > \widehat{J}_{\max}) = o(1)$ uniformly for $h \in \mathcal{H}$. The proof of $\zeta(J) = J$ follows analogously using the condition $p \geq 7d_x/4$. $\qquad\square$

F. U-statistics deviation results

We make use of the following exponential inequality established by Houdré and Reynaud-Bouret [2003], see also Gine and Nickl [2016, Theorem 3.4.8].

Lemma F.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 for all $u > 0$ and $n \in \mathbb{N}$:*

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

where

$$\Lambda_1^2 = \frac{n(n-1)}{2} \mathrm{E}[R^2(Z_1, Z_2)],$$

$$\Lambda_2 = n \sup_{\|\nu\|_{L^2(Z)} \leq 1, \|\kappa\|_{L^2(Z)} \leq 1} \mathrm{E}[R(Z_1, Z_2)\nu(Z_1)\kappa(Z_2)],$$

$$\Lambda_3 = \sqrt{n \sup_z |\mathrm{E}[R^2(Z_1, z)]|},$$

$$\Lambda_4 = \sup_{z_1, z_2} |R(z_1, z_2)|.$$

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 $M_i = \{|Y_i - h_0(X_i)| \leq M_n\}$. Recall that the kernel R_1 is a symmetric function satisfying $\mathrm{E}[R_1(Z,z)] = 0$ for all z.

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

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

$$\Lambda_2 \leq \overline{\sigma}^2 \, n \, s_J^{-2}, \tag{F.2}$$

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

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

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

$$\begin{aligned}
\mathrm{E}_h[R_1^2(Z_1, Z_2)] &\leq \mathrm{E}_h \left| U_1 b^K(W_1)' A' A b^K(W_2) U_2 \right|^2 \\
&= \mathrm{E}_h \left[U^2 b^K(W)' A' A \, \mathrm{E}_h \left[U^2 b^K(W) b^K(W)' \right] A' A b^K(W) \right] \\
&= \mathrm{E}_h \left[(V^J)' \, \mathrm{E}_h \left[V^J (V^J)' \right] V^J \right] = \sum_{j,j'=1}^{J} \mathrm{E}_h[V_j V_{j'}]^2 = v_J^2.
\end{aligned}$$

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

$$\begin{aligned}
\left| \mathrm{E}_h[R_1(Z_1, Z_2)\nu(Z_1)\kappa(Z_2)] \right| &\leq \left| \mathrm{E}_h[U \mathbb{1}_M b^K(W)' \nu(Z)] A' A \, \mathrm{E}_h[U \mathbb{1}_M b^K(W) \kappa(Z)] \right| \\
&\leq \| A \, \mathrm{E}_h[U \mathbb{1}_M b^K(W) \kappa(Z)] \| \, \| A \, \mathrm{E}_h[U \mathbb{1}_M b^K(W) \nu(Z)] \| \\
&\leq \| A G_b^{1/2} \|^2 \sqrt{\mathrm{E} \left[|\, \mathrm{E}_h[U \mathbb{1}_M \kappa(Z)|W] |^2 \right]} \times \sqrt{\mathrm{E} \left[|\, \mathrm{E}_h[U \mathbb{1}_M \nu(Z)|W] |^2 \right]}
\end{aligned}$$

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

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

$$\begin{aligned}
\left| \mathrm{E}_h[R_1^2(Z_1, z)] \right| &\leq \mathrm{E}_h \left| U \mathbb{1}\{|U| \leq M_n\} b^K(W)' A' A b^K(w) u \mathbb{1}\{|u| \leq M_n\} \right|^2 \\
&\leq \| A b^K(w) u \mathbb{1}\{|u| \leq M_n\} \|^2 \, \mathrm{E}_h \| A b^K(W) U \|^2 \\
&\leq \overline{\sigma}^2 M_n^2 \, \zeta_{b,K}^2 \, \| A G_b^{1/2} \|^4,
\end{aligned}$$

again by using Assumption 2(i) and hence the upper bound (F.3) follows.

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

$$\left| R_1(z_1, z_2) \right| \leq \left| u_1 \mathbb{1}\{|u_1| \leq M_n\} b^K(w_1)' A' A b^K(w_2) u_2 \mathbb{1}\{|u_2| \leq M_n\} \right|$$
$$\leq \sup_{u,w} \left\| A b^K(w) u \mathbb{1}\{|u| \leq M_n\} \right\|^2 \leq M_n^2 \zeta_{b,K}^2 \| A G_b^{1/2} \|^2,$$

which completes the proof. $\qquad\square$