cemmap
centre for microdata methods and practice

Adaptive test of conditional moment inequalities

Denis Chetverikov

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP36/12

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Adaptive Test of Conditional Moment Inequalities

By Denis Chetverikov*

Current Version: May, 2012
First Version: November, 2010

Abstract

In this paper, I construct a new test of conditional moment inequalities based on studentized kernel estimates of moment functions. The test automatically adapts to the unknown smoothness of the moment functions, has uniformly correct asymptotic size, and is rate optimal against certain classes of alternatives. Some existing tests have nontrivial power against $n^{-1/2}$-local alternatives of the certain type whereas my method only allows for nontrivial testing against $(n/\log n)^{-1/2}$-local alternatives of this type. There exist, however, large classes of sequences of well-bahaved alternatives against which the test developed in this paper is consistent and those tests are not.

Keywords: Conditional Moment Inequalities, Minimax Rate Optimality.

1 Introduction

Conditional moment inequalities (CMI) are often encountered both in economics and econometrics. In economics, they arise naturally in many models that include behavioral choice, see Pakes (2010) for a survey. In econometrics, they appear in the estimation problems with interval data and problems with censoring, e.g., see Manski and Tamer (2002). In addition, CMI offer a convenient way to study treatment effects in randomized experiments as described in Lee et al. (2011). In the next section, I provide three detailed examples of models with CMI.

*MIT, Department of Economics. Email: dchetver@mit.edu. I thank Victor Chernozhukov for his guidance, numerous discussions and permanent support. I am also grateful to Isaiah Andrews, Jerry Hausman, Kengo Kato, Anton Kolotilin, and Anna Mikusheva for useful comments and discussions. The first version of the paper was presented at the Econometric lunch at MIT on November 18, 2010.

Let $m : \mathbb{R}^d \times \mathbb{R}^k \times \Theta \to \mathbb{R}^p$ be a vector-valued known function. Let (X, W) be a pair of $\mathbb{R}^d$ and $\mathbb{R}^k$-valued random vectors, and $\theta \in \Theta$ a parameter. The CMI can be written as

$$E[m(X, W, \theta)|X] \leq 0 \, a.s. \tag{1.1}$$

I am interested in testing the null hypothesis, H_0, that $\theta = \theta_0$ against the alternative, H_a, that $\theta \neq \theta_0$ based on a random sample $(X_i, W_i)_{i=1}^n$ from the distribution of (X, W). Note that I also allow for conditional moment equalities since they can be written as pairs of the CMI in model (1.1).

Using CMI for inference is difficult because often these inequalities do not identify the parameter. Let

$$\Theta_I = \{\theta \in \Theta : \, E[m(X, W, \theta)|X] \leq 0 \; a.s.\} \tag{1.2}$$

denote the identified set. The model is said to be identified if and only if Θ_I is a singleton. Otherwise, CMI do not identify the parameter θ. For example, nonidentification may happen when the CMI arise from a game-theoretic model with multiple equilibria. Moreover, the parameter may be weakly identified. My approach leads to a test with the correct asymptotic size no matter whether the parameter is identified, weakly identified, or not identified.

Two approaches to robust CMI testing have been developed in the literature. One approach (Andrews and Shi (2010)), is based on converting CMI into an infinite number of unconditional moment inequalities using nonnegative weighting functions. The other approach (Chernozhukov et al. (2009)), is based on estimating moment functions nonparametrically. My method is inspired by the work of Andrews and Shi (2010). To motivate the test developed in this paper, consider two examples of CMI models. These models are highly stylized but convey the main ideas. In the first model, m is multiplicatively separable in θ, i.e. $m(X, W, \theta) = \theta \tilde{m}(X, W)$ for some $\tilde{m} : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ and $\theta \in \mathbb{R}$ with $E[\tilde{m}(X, W)|X] > 0$ a.s. In the second model, m is additively separable in θ, i.e. $m(X, W, \theta) = \tilde{m}(X, W) + \theta$. The identified sets, Θ_I, in these models are $\{\theta \in \mathbb{R} : \theta \leq 0\}$ and $\{\theta \in \mathbb{R} : \theta \leq -\text{ess}\sup_X E[\tilde{m}(X, W)|X]\}$[1] correspondingly. Andrews and Shi (2010) developed a test that has nontrivial power against alternatives of the form $\theta_0 = \theta_{0,n} = C/\sqrt{n}$ for any $C > 0$ in the first model, so their test has extremely high power in this model. It follows from Armstrong (2011a) that their test has low power in the second model, however (e.g., in comparison with the test of Chernozhukov et al. (2009))[2]. In constrast, I construct a test that has

[1] By definition, $\text{ess}\sup_X f(X) = \inf\{M \in \mathbb{R} : \, f(X) \leq M \, a.s.\}$ (essential supremum). If $E[\tilde{m}(X, W)|X]$ is continuous, then essential supremum equals usual supremum.

[2] Andrews and Shi (2010) developed tests based on both Cramer-von Mises and Kolmogorov-Smirnov test statistics. In this paper, I refer to their test with the Kolmogorov-Smirnov test statistic. Most statements are also applicable for Cramer-von Mises test statistic as well, however.

high power in a large class of CMI models including models like that in the second example. At the same time, my test has nearly the same power in models like that described in the first example. The main difference between two approaches is that my test statistic is based on the *studentized* estimates of moments whereas theirs is not. More precisely, Andrews and Shi (2010) consider studentized statistics but modify the variance term so that asymptotic power properties of their test are similar to those of the test with no studentization.

The test of Chernozhukov et al. (2009) also has high power in a large class of CMI models but it requires knowledge of certain smoothness properties of moment functions such as order of differentiability whereas the test developed in this paper does not. Moreover, my test automatically adapts to these smoothness properties selecting the most appropriate weighting function. For this reason, I call the test adaptive. This feature of the test is important because smoothness properties of moment functions are rarely known in practice.

The test statistic in this paper is based on kernel estimates of moment functions $E[m_j(X, W, \theta_0)|X]$ with many bandwidth values using positive kernels[3]. Here $m_j(X, W, \theta)$ denotes j-th component of $m(X, W, \theta)$. I assume that the set of bandwidth values expands as the sample size n increases so that the minimal bandwidth value converges to zero at an appropriate rate while the maximal one is fixed. Since the variance of the kernel estimators varies greatly with the bandwidth value, each estimator is studentized, i.e. it is divided by its estimated standard deviation. The test statistic, $\hat{T}$, is formed as the maximum of these studentized estimates, and large values of $\hat{T}$ suggest that the null hypothesis is violated.

I develop a bootstrap method to simulate a critical value for the test. The method is based on the observation that the distribution of the test statistic is asymptotically independent of the distribution of the noise $\{m(X_i, W_i, \theta_0) - E[m(X_i, W_i, \theta_0)|X_i]\}_{i=1}^n$ apart from its second moment. For reasons similar to those discussed in Chernozhukov et al. (2007) and Andrews and Soares (2010), the distribution of the test statistic in large samples depends heavily on the extent to which the CMI are binding. Moreover, the parameters that measure to what extent the CMI are binding cannot be estimated consistently. I develop a new approach to deal with this problem, which I refer to as the refined moment selection (RMS) procedure. The approach is based on a pretest which is used to decide what counterparts of the test statistic should be used in simulating the critical value for the test. Unlike Andrews and Shi (2010), I use a model-specific, data-driven, critical value for the pretest, which is taken to be a large quantile of the appropriate distribution, whereas they use a deterministic threshold with no reference to the model. I also provide a plug-in critical value for the test. My proof of the bootstrap validity is interesting on its own right because it is unknown whether the test statistic has a limiting distribution.

[3]A kernel is said to be positive if the kernel function is positive on its support.

None of the tests in the literature including mine have power against alternatives in the set Θ_I. Therefore, I consider the alternatives of the form

$$P\{E[m_j(X,W,\theta_0)|X] > 0\} > 0 \,\text{for some}\, j = 1,...,p \tag{1.3}$$

To show that my test has good power properties in a large class of CMI models, I derive its power against alternatives of the form (1.3) assuming that $E[m(X,W,\theta_0)|X]$ is some vector of unrestricted nonparametric functions. In other words, I consider nonparametric classes of alternatives. Once $m(X,W,\theta)$ is specified, it is straightforward to translate my results to the parametric setting. The test developed in this paper is consistent against any fixed alternative outside of the set Θ_I. I also show that my method allows for nontrivial testing against $(n/\log n)^{-1/2}$-local one-dimensional alternatives[4]. Finally, I prove that the test is minimax rate optimal against certain classes of smooth alternatives consisting of moment functions $E[m(X,W,\theta_0)|X]$ that are sufficiently flat at the points of maxima. Minimax rate optimality means that the test is uniformly consistent against alternatives in the mentioned class whose distance from the set of models satisfying (1.1) converges to zero at the fastest possible rate. The requirement that functions should be sufficiently flat cannot be dropped because the test is based on the positive kernels.

The literature concerned with unconditional and conditional moment inequalities is expanding quickly. Published papers on unconditional moment inequalities include Chernozhukov et al. (2007), Romano and Shaikh (2008), Rosen (2008), Andrews and Guggenberger (2009), Andrews and Han (2009), Andrews and Soares (2010), Bugni (2010), Canay (2010), Pakes (2010), and Romano and Shaikh (2010). There is also a large literature on partial identification which is closely related to that on moment inequalities. Methods specific for conditional moment inequalities were developed in Khan and Tamer (2009), Kim (2008), Chernozhukov et al. (2009), Andrews and Shi (2010), Lee et al. (2011), Armstrong (2011a), and Armstrong (2011b). The case of CMI that point identify θ is treated in Khan and Tamer (2009). The test of Kim (2008) is closely related to that of Andrews and Shi (2010). Lee et al. (2011) developed a test based on the minimum distance statistic in the one-sided L_p-norm and kernel estimates of moment functions. The advantage of their approach comes from simplicity of their critical value for the test, which is an appropriate quantile of the standard Gaussian distribution. Their test is not adaptive, however, since only one bandwidth value is used. Armstrong (2011a) developed a new method for computing the critical value for the test statistic of Andrews and Shi (2010) that leads to a more powerful test than theirs but the

[4]In this paper, the term 'local one-dimensional alternative' is used to refer to a sequence of models $m = m_n(X,W,\theta_0)$ such that $E[m_n(X,W,\theta_0)|X] = a_n f(X)$ for some sequence of positive numbers $\{a_n\}_{n=1}^{\infty}$ converging to zero where $f : \mathbb{R}^d \to \mathbb{R}^p$ satisfies $P\{f_j(X) > 0\} > 0$ for some $j = 1,...,p$.

resulting test is not robust. In particular, his method cannot be used in CMI models like that described in the first example above. Armstrong (2011b), which was written independently and at the same time as this paper, considered a test statistic similar to that used in this paper and derived a critical value such that the whole identified set is contained in the confidence region with probability approaching one. In other words, he focused on estimation rather than inference.

Finally, an important related paper in the statistical literature is Dumbgen and Spokoiny (2001). They consider testing qualitative hypotheses in the ideal Gaussian white noise model where a researcher observes a stochastic process that can be represented as a sum of the mean function and a Brownian motion. In particular, they developed a test of the hypothesis that the mean function is (weakly) negative almost everywhere. Though their test statistic is somewhat related to that used in this paper, the technical details of their analysis are quite different.

The rest of the paper is organized as follows. The next section discusses some examples of CMI models. Section 3 formally introduces the test. The main results of the paper are presented in section 4. Extensions to the cases of infinitely many CMI and local CMI are provided in section 5. A Monte Carlo simulation study is described in section 6. There I provide an example of an alternative with a well-behaved moment function such that the test developed in this paper rejects the null hypothesis with probability higher than 80% while the rejection probability of all competing tests does not exceed 20%. Brief conclusions are drawn in section 7. Finally, all proofs are contained in the Appendix.

2 Examples

In this section, I provide three examples of CMI models.

Incomplete Models of English Auctions. My first example follows Haile and Tamer (2003) treatment of English auctions under weak conditions. The popular model of English auctions suggested by Milgrom and Weber (1982) assumes that each bidder is holding down the button while the price for the object is going up continuously until she wants to drop out. The price at the moment of dropping out is her bid. It is well-known that the dominant strategy in this model is to make a bid equal to her valuation of the object. In practice, participants usually call out bids, however. Hence, the price rises in jumps, and the bid may not be equal to person's valuation of the object. In this situation, the relation between bids and valuations of the object depends crucially on the modeling assumptions. Haile and Tamer (2003) derived certain bounds on the distribution function of valuations based on minimal assumptions of rationality.

Suppose that we have an auction with m bidders whose valuations of the object are drawn independently from the distrubution $F(\cdot, X)$ where X denotes observable characterics of the object. Let $b_1, ..., b_m$ denote highest bids of each bidder. Let $b_{1:m} \leq ... \leq b_{m:m}$ denote the ordered sequence of bids $b_1, ..., b_m$. Assuming that bids do not exceed bidders' valuations, Haile and Tamer (2003) derived the following upper bound on $F(\cdot, X)$:

$$E[\phi^{-1}(F(v, X)) - I\{b_{i:m} \leq v\}|X] \leq 0 \, a.s. \tag{2.1}$$

for all $v \in \mathbb{R}$ and $i = 1, ..., m$ where $\phi(\cdot)$ is a certain (known) increasing function, see equation (3) in Haile and Tamer (2003). A similar lower bound follows from the assumption that bidders do not allow opponents to win at a price they would like to beat. Parameterizing the function $F(\cdot, \cdot)$ and selecting a finite set $V = \{v_1, ..., v_p\}$ gives inequalities of the form (1.1).

Interval Data. In some cases, especially when data involve personal information like individual income or wealth, one has to deal with interval data. Suppose we have a mean regression model

$$Y = f(X, V) + \varepsilon \tag{2.2}$$

where $E[\varepsilon|X, V] = 0$ a.s. and V is a scalar random variable. Suppose that we observe X and Y but do not observe V. Instead, we observe V_0 and V_1, called brackets, such that $V \in [V_0, V_1]$ a.s. In empirical analysis, brackets may arise because a respondent refuses to provide information on V but provides an interval to which V belongs. Following Manski and Tamer (2002), assume that $f(X, V)$ is weakly increasing in V and $E[Y|X, V] = E[Y|X, V, V_0, V_1]$. Then it is easy to see that

$$E[I\{V_1 \leq v\}(Y - f(X, v))|X, V_0, V_1] \leq 0 \tag{2.3}$$

and

$$E[I\{V_0 \geq v\}(Y - f(X, v))|X, V_0, V_1] \geq 0 \tag{2.4}$$

for all $v \in \mathbb{R}$. Again, parameterizing the function $f(\cdot, \cdot)$ and selecting a finite set $V = \{v_1, ..., v_p\}$ gives inequalities of the form (1.1).

Treatment Effects. Suppose that we have a randomized experiment where one group of people gets a new treatment while the control group gets a placebo. Let $D = 1$ if the person gets the treatment and 0 otherwise. Let p denote the probability that $D = 1$. Let X denote person's observable characteristics and Y denote the realized outcome. Finally, let Y_0 and Y_1 denote the counterfactual outcomes had the person received a placebo or the new medicine respectively. Then $Y = DY_1 + (1 - D)Y_0$. The question of interest is whether the new medicine has a positive expected impact uniformly over all possible charactersics X. In

other words, the null hypothesis, H_0, is that

$$E[Y_1 - Y_0|X] \geq 0 \, a.s. \tag{2.5}$$

Since in randomized experiments D is independent of X, Lee et al. (2011) showed that

$$E[Y_1 - Y_0|X] = E[DY/p - (1-D)Y/(1-p)|X] \tag{2.6}$$

Combining (2.5) and (2.6) gives CMI.

3 The Test

In this section, I present the test statistic and give two bootstrap methods to simulate critical values. The analysis in this paper is conducted conditional on the set of values $\{X_i\}_{i=1}^{\infty}$, so all probabilistic statements excluding those in lemmas 3 and 4 in the Appendix should be understood conditional on $\{X_i\}_{i=1}^{\infty}$ for almost all sequences $\{X_i\}_{i=1}^{\infty}$. Lemmas 3 and 4 provide certain conditions that ensure that the assumptions used in this paper hold for almost all sequences $\{X_i\}_{i=1}^{\infty}$.

For fixed θ_0, let $f(X) = E[m(X, W, \theta_0)|X]$. Then under the null hypothesis,

$$f(X) \leq 0 \, a.s. \tag{3.1}$$

In addition, let $Y_i = m(X_i, W_i, \theta_0)$ and $\varepsilon_i = Y_i - f(X_i)$ so that $E[\varepsilon_i|X_i] = 0$ a.s. $(i = 1, ..., n)$. Finally, let $f_1, ..., f_p$ denote components of f.

Section 3.1 defines the test statistic assuming that $\Sigma_i = E[\varepsilon_i \varepsilon_i^T|X_i]$ is known for each $i = 1, ..., n$. Section 3.2 gives two bootstrap methods to simulate critical values. The first one is based on plug-in asymptotics, while the second one uses the refined moment selection (RMS) procedure. Section 3.2 also provides some intuition of why these procedures lead to the correct asymptotic size of the test. When Σ_i is unknown, it should be estimated from the data. Section 3.3 shows how to construct an appropriate estimator $\hat{\Sigma}_i$ of Σ_i. The feasible version of the test will be based on substituting $\hat{\Sigma}_i$ for Σ_i both in the test statistic and in the critical value.

3.1 The Test Statistic

The test statistic in this paper is based on a kernel estimator of the vector-valued function f. Let $K : \mathbb{R}^d \to \mathbb{R}_+$ be some kernel. For bandwidth value $h \in \mathbb{R}_+$, let $K_h(x) = K(x/h)/h^d$.

For each pair of observations $i, j = 1, ..., n$, denote the weight function

$$w_h(X_i, X_j) = \frac{K_h(X_i - X_j)}{\sum_{k=1}^n K_h(X_i - X_k)} \tag{3.2}$$

Then the kernel estimator of $f_m(X_i)$ is

$$\hat{f}_{(i,m,h)} = \sum_{j=1}^n w_h(X_i, X_j) Y_{j,m} \tag{3.3}$$

where $Y_{j,m}$ denotes m-th component of Y_j[5]. Conditional on $\{X_i\}_{i=1}^n$, the variance of the kernel estimator $\hat{f}_{(i,m,h)}$ is

$$V^2_{(i,m,h)} = \sum_{j=1}^n w_h^2(X_i, X_j) \Sigma_{j,mm} \tag{3.4}$$

where $\Sigma_{j,m_1 m_2}$ denotes (m_1, m_2) component of Σ_j.

Next, consider a finite set of bandwidth values $H = \{h = h_{\max} a^k : h \geq h_{\min}, k = 0, 1, 2, ...\}$ for some $h_{\max} > h_{\min}$ and $a \in (0, 1)$. For simplicity, I assume that $h_{\min} = h_{\max} a^k$ for some $k \in \mathbb{N}$ so that $h_{\min}$ is included in H. I assume that as the sample size n increases, $h_{\min}$ converges to zero while $h_{\max}$ is fixed. For practical purposes, I recommend setting $h_{\max} = \max_{i,j=1,...,n} \|X_i - X_j\|/2$, $h_{\min} = h_{\max}(0.04/n^{0.95})^{1/(3d)}$, and $a = 0.8$. This choice of parameters is consistent with the theory presented in the paper and also worked well in my simulations. Note that $h_{\min}$ is chosen so that the kernel estimator uses on average roughtly 15 data points when $n = 250$.

For each bandwidth value $h \in H$, choose a subset I_h of observations such that $\|X_j - X_k\| > 2h$ for all $j, k \in I_h$ with $j \neq k$ and for each $i = 1, ..., n$, there exists an element $j(i) \in I_h$ such that $\|X_i - X_{j(i)}\| \leq 2h$ where $\|\cdot\|$ denotes the Eucledian norm on $\mathbb{R}^d$. I refer to I_h as a set of test points. The choice of I_h may be random, but it is important to select I_h independently of $\{Y_i\}_{i=1}^n$. Conditional on $\{X_i\}_{i=1}^n$, I assume that I_h is nonstochastic. It will be assumed in the next section that $K(x) = 0$ for any $x \in \mathbb{R}^d$ such that $\|x\| > 1$. Thus, random variables $\{\hat{f}_{(i,m,h)}\}_{i \in I_h}$ are jointly independent for any fixed $m = 1, ..., p$ and $h \in H$ conditional on $\{X_i\}_{i=1}^n$. Finally, denote $S = \{(i, m, h) : i \in I_h, m = 1, ..., p, h \in H\}$.

Based on this notation, the test statistic is

$$T = \max_{s \in S} \frac{\hat{f}_s}{V_s} \tag{3.5}$$

Let me now explain why the optimal bandwidth value depends on the smoothness proper-

[5]The estimator of $f_m(X_i)$ is usually denoted by $\hat{f}_m(X_i)$. I use nonstandard notation $\hat{f}_{(i,m,h)}$ because it will be more convenient later in the paper.

ties of the components $f_1, ..., f_p$ of f. Without loss of generality, consider $j = 1$. Suppose that $f_1(X)$ is flat. Then $f_1(X)$ is positive on the large subset of its domain whenever its maximal value is positive. Hence, the maximum of $\hat{T}$ will correspond to a large bandwidth value because the variance of the kernel estimator, which enters the denominator of the test statistic, decreases with the bandwidth value. On the other hand, if $f_1(X)$ is allowed to have peaks, then there may not exist a large subset where it is positive. Hence, large bandwidth values may not yield large values of $\hat{T}$, and small bandwidth values should be used. I circumvent the problem of bandwidth selection by considering the set of bandwidth values jointly, and let the data determine the best bandwidth value. In this sense, my test adapts to the smoothness properties of $f(X)$. This allows me to construct a test with good uniform power properties over possible degrees of smoothness for $f(X)$.

When Σ_i is unknown, which is usually the case in practice, one can define $\hat{V}^2_{(i,m,h)} = \sum_{j=1}^n w_h^2(X_i, X_j)\hat{\Sigma}_{j,mm}$ and use

$$\hat{T} = \max_{s \in S} \frac{\hat{f}_s}{\hat{V}_s} \tag{3.6}$$

instead of T, where $\hat{\Sigma}_j$ is some estimator of Σ_j. Some possible estimators are discussed in section 3.3.

3.2 Critical Values

Suppose we want to construct a test of size α. This subsection explains how to simulate a critical value $c_{1-\alpha}$ for the test statistic $\hat{T}$ based on two bootstrap methods. One method is based on plug-in asymptotics while the other one uses the refined moment selection (RMS) procedure. The resulting test will reject the null hypothesis if and only if $\hat{T} > c_{1-\alpha}$.

The first method relies on two observations. First, it is easy to see that, for a fixed distribution of disturbances $\{\varepsilon_i\}_{i=1}^n$, the maximum of $1-\alpha$ quantile of the test statistic $\hat{T}$ over all possible functions f satisfying $f \leq 0$ a.s. corresponds to $f = 0_p$. Second, lemmas 9 and 11 in the Appendix show that the distribution of the statistic $\hat{T}$ is asymptotically independent of the distrubution of disturbances $\{\varepsilon_i : i = 1, ..., n\}$ apart from their second moments $\{\Sigma_i : i = 1, ..., n\}$. These observations suggest that one can simulate $c_{1-\alpha}$ (denoted by $c^{PIA}_{1-\alpha}$) by the following procedure:

1. For each $i = 1, ..., n$, simulate $\tilde{Y}_i \sim N(0_p, \hat{\Sigma}_i)$ independently across i.
2. Calculate $T^{PIA} = \max_{(i,m,h) \in S} \sum_{j=1}^n w_h(X_i, X_j)\tilde{Y}_{j,m}/\hat{V}_{(i,m,h)}$.
3. Repeat steps 1 and 2 independently B times for some large B to obtain $\{T^{PIA}_b : b = 1, ..., B\}$.

4. Let $c_{1-\alpha}^{PIA}$ be $1-\alpha$th empirical quantile of $\{T_b^{PIA}\}_{b=1}^B$.

The second method is based on the refined moment selection (RMS) procedure. It gives a more powerful test while maintaining the required size. The method relies on the observation that $|\hat{T}| = O_p(\sqrt{\log n})$ if $f = 0_p$ (see lemmas 8, 9, and 11 in the Appendix) while $\hat{f}_{i,m,h}/\hat{V}_{(i,m,h)} \to -\infty$ at a polynomial rate if $f_m(X) < 0$ for X satisfying $\|X - X_i\| < h$. Such terms will have asymptotically negligible effect on the distribution of $\hat{T}$, so we can ignore corresponding terms in the simulated statistic. Therefore, one can simulate $c_{1-\alpha}$ (denoted by $c_{1-\alpha}^{RMS}$) as follows. First, let $\gamma < \alpha/2$ be some small positive number (truncation parameter). Second, use the plug-in bootstrap to find $c_{1-\gamma}^{PIA}$. Denote

$$S^{RMS} = \{s \in S : \hat{f}_s/\hat{V}_s > -2c_{1-\gamma}^{PIA}\} \tag{3.7}$$

Third, run the following procedure:

1. For each $i = 1, ..., n$, simulate $\tilde{Y}_i \sim N(0_p, \hat{\Sigma}_i)$ independently across i.
2. Calculate $T^{RMS} = \max_{(i,m,h)\in S^{RMS}} \sum_{j=1}^n w_h(X_i, X_j)\tilde{Y}_{j,m}/\hat{V}_{(i,m,h)}$.
3. Repeat steps 1 and 2 independently B times for some large B to obtain $\{T_b^{RMS} : b = 1, ..., B\}$.
4. Let $c_{1-\alpha}^{RMS}$ be $1-\alpha$th empirical quantile of $\{T_b^{RMS}\}_{b=1}^B$.

In the next section, it will be assumed that $\gamma = \gamma_n \to 0$ as $n \to 0$. So, I recommend setting γ as a small fraction of α, for example $\gamma = 0.01$ for $\alpha = 0.05$. Alternatively, one can set $\gamma = 0.1/\log(n)$, similar to Chernozhukov et al. (2009)[6].

3.3 Estimating Σ_i

Let me now explain how one can estimate Σ_i. The literature on estimating Σ_i is huge. Among other papers, it includes Rice (1984), Muller and Stadtmuller (1987), Hardle and Tsybakov (1997), and Fan and Yao (1998). For scalar-valued Y_i, available estimators are described in Horowitz and Spokoiny (2001). All those estimators can be immediately generalized to vector-valued Y_i. For concreteness, I describe one estimator here. Choose a bandwidth value $b_n > 0$. For $i = 1, ..., n$, let $J(i) = \{j = 1, ..., n : \|X_j - X_i\| \le b_n\}$. If $J(i)$ has an odd number of elements, drop one arbitrarily selected observation. Partition $J(i)$ into pairs using a map $k : J(i) \to J(i)$ satisfying $k(j) \neq j$ and $k(k(j)) = j$ for all $j \in J(i)$. Let $|J(i)|$ denote the

[6]Note also that if γ is comparable with α, one can do a finite sample adjustment of the critical value by taking $1 - \alpha + 2\gamma$th quantile of $\{T_b^{RMS}\}_{b=1}^B$ at step 4 of the procedure above.

number of elements in $J(i)$. Then Σ_i can be estimated by

$$\hat{\Sigma}_i = \sum_{j \in J(i)} (Y_{k(j)} - Y_j)(Y_{k(j)} - Y_j)^T / (2|J(i)|) \tag{3.8}$$

Lemma 1 in the Appendix gives certain conditions that ensure that this estimator will be uniformly consistent for Σ_i over $i = 1, ..., n$ with a polynomial rate, i.e.

$$\max_{i=1,...,n} \|\hat{\Sigma}_i - \Sigma_i\|_o = o_p(n^{-\kappa}) \tag{3.9}$$

for some $\kappa > 0$ where $\|\cdot\|_o$ denotes the spectral norm on the space of $p \times p$-dimensional symmetric matrices corresponding to the Eucledian norm on $\mathbb{R}^p$. To choose the bandwidth value b_n in practice, one can use appropriately modified cross validation. An advantage of this estimator is that it is fully adaptive with respect to the smoothness properties of f.

The intuition behind this estimator is based on the following argument. Note that $k(j)$ is chosen so that $X_{k(j)}$ is close to X_j. If the function f is continuous,

$$Y_{k(j)} - Y_j = f(X_{k(j)}) - f(X_j) + \varepsilon_{k(j)} - \varepsilon_j \approx \varepsilon_{k(j)} - \varepsilon_j \tag{3.10}$$

so that

$$E[(Y_{k(j)} - Y_j)(Y_{k(j)} - Y_j)^T | \{X_i\}_{i=1}^n] \approx \Sigma_{k(j)} + \Sigma_j \tag{3.11}$$

since $\varepsilon_{k(j)}$ is independent of ε_j. If b_n is small enough and $\Sigma(X)$ is continuous, $\Sigma_{k(j)} + \Sigma_j \approx 2\Sigma_i$ since $\|X_{k(j)} - X_i\| \leq b_n$ and $\|X_j - X_i\| \leq b_n$.

4 The Main Results

This section presents my main results. Section 4.1 gives regularity conditions. Section 4.2 describes size properties of the test. Section 4.3 explains the behavior of the test under a fixed alternative. Section 4.4 derives the rate of consistency of the test against local one-dimensional alternatives mentioned in the introduction. Section 4.5 shows the rate of uniform consistency against certain classes of smooth alternatives. Section 4.6 presents the minimax rate-optimality result.

4.1 Assumptions

Let C_j $(j = 1, ..., 6)$ be a set of strictly positive and finite constants independent of the sample size n. Let $M_h(X_i)$ be the number of elements in the set $\{X_j : \|X_j - X_i\| \leq h,\ j = 1, ..., n\}$. Results in this paper will be proven under the following assumptions.

Assumption 1. *(i) Design points* $\{X_i\}_{i=1}^{\infty}$ *are nonstochastic. (ii)* $C_1 nh^d \leq M_h(X_i) \leq C_2 nh^d$ *for all* $i \in \mathbb{N}$ *and* $h \in H = H_n$.

The design points are nonstochastic because the analysis is conducted conditional on $\{X_i\}_{i=1}^{\infty}$. In addition, assumption 1 states that the number of design points in certain neighborhoods of each design point is proportional to the volume of the neighborhood with the coefficient of proportionality bounded from above and away from zero. It is stated in Horowitz and Spokoiny (2001) that assumption 1 holds in an iid setting with probability approaching one as the sample size increases if the distribution of X_i is absolutely continuous with respect to Lebegue measure, has bounded support, and has density bounded away from zero on the support. This statement is actually wrong unless one makes some extra assumptions. Lemma 3 in the Appendix gives a counter-example. Instead, lemma 4 shows that assumption 1 holds for large n a.s. if, in addition, I assume that the density of X_i is bounded from above, and that the support of X_i is a convex set. Necessity of the density boundedness is obvious. Convexity of the support is not necessary for assumption 1 but it strikes a good balance between generality and simplicity. In general, one must deal with some smoothness properties of the boundary of the support. Note that the statement "for large n a.s." is stronger than "with probability approaching one". Note also that assumption 1(ii) requires inequalities to hold for all $i \in \mathbb{N}$, not just for $i = 1, ..., n$.

Assumption 2. *(i) Disturbances* $\{\varepsilon_i\}_{i=1}^{\infty}$ *are independent* $\mathbb{R}^p$*-valued random variables with* $\mathbb{E}[\varepsilon_i] = 0$ *for all* $i = 1, ..., \infty$. *(ii)* $E[\|\varepsilon_i\|^3] \leq C_3$ *for all* $i = 1, ..., \infty$. *(iii)* $\Sigma_{i,mm} \geq C_4$ *for all* $i = 1, ..., \infty$ *and* $m = 1, ..., p$.

Finite third moment of disturbances is used in the derivation of a certain invariance principle with the rate of convergence. As in the classical central limit theorem, two finite moments are sufficient to prove weak convergence but more finite moments are necessary if we are interested in the rate of convergence. I assume that the variance of each component of disturbances is bounded away from zero for simplicity of the presentation. Since I use studentized kernel estimates, without this assumption, it would be necessary to truncate the variance of the kernel estimators from below with truncation level slowly converging to zero. That would complicate the derivation of the main results without changing the main ideas.

Before stating assumption 3, let me give formal definitions of Holder smoothness class $\mathcal{F}(\tau, L)$ and its subsets $\mathcal{F}_\varsigma(\tau, L)$. For d-tuple of nonnegative integers $\alpha = (\alpha_1, ..., \alpha_d)$ with $|\alpha| = \alpha_1 + ... + \alpha_d$, function $g : \mathbb{R}^d \to \mathbb{R}$, and $x = (x_1, ..., x_d) \in \mathbb{R}^d$, denote

$$D^\alpha g(x) = \frac{\partial^{|\alpha|} g}{\partial x_1^{\alpha_1} ... \partial x_d^{\alpha_d}}(x) \tag{4.1}$$

whenever it exists. For $\tau > 0$ and $L > 0$, it is said that the function $g : \mathbb{R}^d \to \mathbb{R}$ belongs to the class $\mathcal{F}(\tau, L)$ if (i) g has continuous partial derivatives up to order $[\tau]$, (ii) for any $\alpha = (\alpha_1, ..., \alpha_d)$ such that $|\alpha| = [\tau]$ and $x, y \in \mathbb{R}^d$,

$$|D^\alpha g(x) - D^\alpha g(y)| \leq L\|x - y\|^{\tau - [\tau]} \tag{4.2}$$

and (iii) for any $\alpha = (\alpha_1, ..., \alpha_d)$ such that $|\alpha| \leq [\tau]$ and any $x \in \mathbb{R}^d$,

$$|D^\alpha g(x)| \leq L \tag{4.3}$$

Here $[\tau]$ denotes the largest integer strictly smaller than τ. Let $S^{d-1} = \{l \in \mathbb{R}^d : \|l\| = 1\}$ denote the space of directions in $\mathbb{R}^d$. For any $g \in \mathcal{F}(\tau, L)$, $x = (x_1, ..., x_d) \in \mathbb{R}^d$, and $l \in S^{d-1}$, let $g^{(k,l)}(x)$ denote k-th derivative of function g in direction l at point x whenever it exists[7]. For $\varsigma = 1, ..., [\tau]$, let $\mathcal{F}_\varsigma(\tau, L)$ denote the class of all elements of $\mathcal{F}(\tau, L)$ such that for any $g \in \mathcal{F}_\varsigma(\tau, L)$ and $l \in S^{d-1}$, $g^{(k,l)}(x) = 0$ for all $k = 1, ..., \varsigma$ whenever $g^{(1,l)}(x) = 0$, and there exist $x = (x_1, ..., x_d) \in \mathbb{R}^d$ and $l \in S^{d-1}$ such that $g^{(\varsigma+1,l)}(x) \neq 0$ and $g^{(1,l)}(x) = 0$. If $\tau \leq 1$, I set $\varsigma = 0$ and $\mathcal{F}_\varsigma(\tau, L) = \mathcal{F}(\tau, L)$.

Assumption 3. *$f_m \in \mathcal{F}_\varsigma(\tau, L)$ for all $m = 1, ..., p$ for some $\tau > 0$, $L > 0$, and $\varsigma = 1, ..., [\tau]$.*

For simplicity of notation, I assume that all components of f have the same smoothness properties. This assumption is used in the derivation of the power properties of the test.

Assumption 4. *(i) The set of bandwidth values has the following form: $H = H_n = \{h = h_{\max} a^k : h \geq h_{\min}, k = 0, 1, 2, ...\}$ where $a \in (0, 1)$, $h_{\max} = \max_{i,j=1,...,n} \|X_i - X_j\|/2$ and $h_{\min} = h_{\min,n} \to 0$ as $n \to \infty$. (ii) For some $\epsilon > 0$, $n^{1-\epsilon} h_{\min}^{3d} > C_5$ for all n.*

According to this assumption, the maximal bandwidth value, $h_{\max}$, is independent of n. Its value is chosen to match the radius of the support of design points. It is intented to detect deviations from the null hypothesis in the form of flat alternatives. The minimal bandwidth value, $h_{\min}$, converges to zero as the sample size increases in such a way that the number of bandwidth values in the set H_n is growing at a logarithmic rate or slower. The minimal bandwidth value is intended to detect alternatives with narrow peaks. Assumption 4(ii) is a key condition used to establish an invariance principle that shows that the distribution of $\hat{T}$ asymptotically depends on the distribution of disturbances $\{\varepsilon_i\}$ only through their covariances $\{\Sigma_i\}$.

Assumption 5. *The truncation parameter, γ, satisfies $\gamma = \gamma_n \to 0$ as $n \to \infty$.*

[7]Let $w : \mathbb{R} \to \mathbb{R}$ be given by $w(t) = g(x + tl)$. By definition, $g^{(k,l)}(x) = w^{(k)}(0)$.

This assumption is used in the proof that the test is asymptotically not conservative.

Assumption 6. *Estimators* $\hat{\Sigma}_i$ *of* Σ_i *satisfy* $\max_{i=1,...,n} \|\hat{\Sigma}_i - \Sigma_i\|_o = o_p(n^{-\kappa})$ *with some* $\kappa > 0$ *where* $\|\cdot\|_o$ *denotes the spectral norm on the space of* $p \times p$*-dimensional symmetric matrices corresponding to the Euclidean norm on* $\mathbb{R}^p$*.*

Assumption 6 is satisfied for $\hat{\Sigma}_i$ described in section 3.3. In practice, due to the curse of dimensionality, it might be useful to use some parametric or semi-parametric estimators of Σ_i instead of the estimator described in section 3.3. For example, if we assume that $\Sigma_i = \Sigma_j$ for all $i, j = 1, ..., n$, then the estimator of Rice (1984) (or its multivariate generalization) is $1/\sqrt{n}$-consistent. In this case, assumption 6 will be satisfied with $\kappa = 1/2 - \phi$ for arbitrarily small $\phi > 0$.

Assumption 7. *(i) The kernel* K *is positive and supported on* $\{x \in \mathbb{R}^d : \|x\| \leq 1\}$*. (ii)* $K(x) \leq 1$ *for all* $x \in \mathbb{R}^d$ *and* $K(x) \geq C_6$ *for all* $\|x\| \leq 1/2$*.*

I assume that the kernel function is positive on its support. Many kernels satisfy this assumption. For example, one can use rectangular, triangular, parabolic, or biweight kernels. See Tsybakov (2009) for the definitions. On the other hand, the requirement that the kernel is positive on its support excludes higher-order kernels, which are necessary to achieve minimax optimal testing rate over large classes of smooth alternatives. I require positive kernels because of their negativity-invariance property, which means that any kernel smoother with a positive kernel maps the space of negative functions into itself. This property is essential for obtaining a test with the correct asymptotic size when smoothness properties of moment functions are unknown. With higher-order kernels, one has to assume undersmoothing so that the bias of the estimator is asymptotically negligible in comparison with its standard deviation. Otherwise, large values of $\hat{T}$ might be caused by large values of the bias term relative to the standard deviation of the estimator even though all components of $f(X)$ are negative. However, for undersmoothing, one has to know the smoothness properties of $f(X)$. In constrast, with positive kernels, the set of bandwidth values can be chosen without reference to these smoothness properties. In particular, the largest bandwidth value can be chosen to be independent of the sample size n. Nevertheless, the test developed in this paper will be rate optimal in the minimax sense against class $\mathcal{F}_{[\tau]}(\tau, L)$ when $\tau > d$.

Assumption 8. *(i) For every* $h \in H_n$*, the set of test points* $I_h = I_{h,n}$ *is such that* $\|X_i - X_j\| > 2h$ *for all* $i, j \in I_{h,n}$ *with* $i \neq j$ *and for each* $i = 1, ..., n$*, there exists an element* $j(i) \in I_{h,n}$ *such that* $\|X_i - X_{j(i)}\| \leq 2h$*. (ii)* $S = S_n = \{(i, m, h) : i \in I_{h,n}, m = 1, ..., p, h \in H_n\}$*.*

Assumptions 1-3 concern with the data-generating process (model) while assumptions 4-8 deal with the test. The asymptotic results in this paper will be shown to hold uniformly

over all data-generating processes satisfying assumptions 1-3. For that purpose, the following notation will be useful. Let $w = (f, \{\varepsilon_i\}_{i=1}^{\infty}, \{X_i\}_{i=1}^{\infty})$ denote the data-generating process. Here, $\{\varepsilon_i\}_{i=1}^{\infty}$ is a sequence of random vectors, and $\{X_i\}_{i=1}^{\infty}$ is a sequence of nonstochastic points. Let $\mathcal{G}$ denote the set of all triples w satisfying assumptions 1-3. For every model $w \in \mathcal{G}$, let $E_w[\cdot]$ denote the expectation calculated assuming the data-generating process w.

4.2 Size Properties of the Test

Analysis of size properties of the test is complicated because it is unknown whether the test statistic has a limiting distribution. Instead, I use a finite sample approach based on the Lindeberg method. For each sample size n, this method gives an upper error bound on approximating the expectation of smooth functionals of the test statistic by the expectation calculated assuming Gaussian noise $\{\varepsilon_i\}_{i=1}^{n}$. I also derive a simple lower bound on the growth rate of the pdf of the test statistic to show that the expectation of smooth functionals can be used to approximate the expectation of indicator functions. Combining these results leads to the approximation of the cdf of the test statistic by its cdf calculated assuming Gaussian disturbances with an explicit error bound. This allows me to derive certain conditions which ensure that the error converges to zero as the sample size n increases, which is a key step in establishing the bootstrap validity.

Let $\mathcal{G}_0$ and $\mathcal{G}_{00}$ denote the set of all elements $(f, \{\varepsilon_i\}_{i=1}^{\infty}, \{X_i\}_{i=1}^{\infty})$ of $\mathcal{G}$ satisfying $f \leq 0$ a.s. and $f = 0$ a.s. correspondingly. The first theorem states that the test has correct asymptotic size uniformly over the class of models $\mathcal{G}_0$ both for plug-in and RMS critical values. In addition, the test is nonconservative as the size of the test converges to the required level α uniformly over the class of models $\mathcal{G}_{00}$.

Theorem 1. *Let assumptions 4-8 hold. Then for $P = PIA$ or RMS,*

$$\inf_{w \in \mathcal{G}_0} P_w\{\hat{T} \leq c_{1-\alpha}^P\} \geq 1 - \alpha + o(1) \tag{4.4}$$

In addition,

$$\sup_{w \in \mathcal{G}_{00}} P_w\{\hat{T} \leq c_{1-\alpha}^P\} = 1 - \alpha + o(1) \tag{4.5}$$

Remark 1. (i) Note that assumptions 1-3 are implicitly imposed in the theorem since $w \in \mathcal{G}_0$.

(ii) Proofs of all results are presented in the Appendix.

4.3 Consistency Against a Fixed Alternative

Let me introduce a distance between the model $w = (f, \{\varepsilon_i\}_{i=1}^{\infty}, \{X_i\}_{i=1}^{\infty}) \in \mathcal{G}$ and the null hypothesis:

$$\rho(w, H_0) = \sup_{i=1,\ldots,\infty;\, m=1,\ldots,p} [f_m(X_i)]_+ \tag{4.6}$$

For any alternative outside of the set Θ_I, $\rho(w, H_0) > 0$. In this section, I argue that the test is consistent against any fixed alternative $w \in \mathcal{G}$ with $\rho(w, H_0) > 0$. Moreover, I show that the test is consistent uniformly against alternatives whose distance from the null hypothesis is bounded away from zero. For $\rho > 0$, let $\mathcal{G}_\rho$ denote the subset of all elements w of $\mathcal{G}$ such that $\rho(w, H_0) \geq \rho$. Then

Theorem 2. *Let assumptions 4-8 hold. Then for $P = PIA$ or RMS,*

$$\sup_{w \in \mathcal{G}_\rho} P_w\{\hat{T} \leq c_{1-\alpha}^P\} \to 0 \tag{4.7}$$

as $n \to \infty$.

4.4 Consistency Against Local One-Dimensional Alternatives

Let $w(0) = (f^0, \{\varepsilon_i\}_{i=1}^{\infty}, \{X_i\}_{i=1}^{\infty}) \in \mathcal{G}$ be such that $\rho(w(0), H_0) > 0$. For some sequence $\{a_n\}_{n=1}^{\infty}$ of positive numbers converging to zero, denote $f^n = a_n f^0$, and let $w_n = (f^n, \{\varepsilon_i\}_{i=1}^{\infty}, \{X_i\}_{i=1}^{\infty})$ be a sequence of local alternatives. I refer to such sequences as local one-dimensional alternatives. This section establishes the consistency of the test against such alternatives whenever $\sqrt{n/\log n} a_n \to \infty$.

Theorem 3. *Let assumptions 4-8 hold. Then for $P = PIA$ or RMS,*

$$P_{w_n}\{\hat{T} \leq c_{1-\alpha}^P\} \to 0 \tag{4.8}$$

as $n \to \infty$ if $\sqrt{n/\log n} a_n \to \infty$.

Remark 2. Recall the CMI model from the first example mentioned in the introduction where $m(X, W, \theta) = \theta \tilde{m}(X, W)$ and $E[\tilde{m}(X, W)|X] > 0$ a.s. The theorem above shows that the test developed in this paper is consistent against sequences of alternatives $\theta_0 = \theta_{0,n}$ whenever $\sqrt{n/\log n}\theta_{0,n} \to \infty$ in this model whereas the test of Andrews and Shi (2010) is consistent whenever $\sqrt{n}\theta_{0,n} \to \infty$. Hence, my test is consistent against nearly the same sequence of alternatives in this model as the test of Andrews and Shi (2010). The additional $\sqrt{\log n}$ factor is the cost for having higher power in other classes of models.

4.5 Uniform Consistency Against Holder Smoothness Classes

In this section, I present the rate of uniform consistency of the test against the class $\mathcal{F}_\varsigma(\tau, L)$ under certain additional constraints. These additional constraints are needed to deal with some boundary effects. Let $S = \text{cl}\{X_i : i \in \mathbb{N}\}$ denote the closure of the infinite set of design points. For any $\vartheta > 0$, let S_ϑ be the subset of S such that for any $x \in S_\vartheta$, the ball with center at x and radius ϑ, $B_\vartheta(x)$, is contained in S, i.e. $B_\vartheta(x) \subset S$. Denote $\zeta = \min(\varsigma + 1, \tau)$. When $\zeta \leq d$, set $\vartheta = \vartheta_n = 4\sqrt{d}h_{\min}$. When $\zeta > d$, set $\vartheta = \vartheta_n = 4\sqrt{d}(\log n/n)^{1/(2\zeta+d)}$. Let $\mathbb{N}_\vartheta = \{i \in \mathbb{N} : X_i \in S_\vartheta\}$. For any $w = (f, \{\varepsilon_i\}_{i=1}^\infty, \{X_i\}_{i=1}^\infty) \in \mathcal{G}$, let

$$\rho_\vartheta(w, H_0) = \sup_{i \in \mathbb{N}_\vartheta,\, m=1,\ldots,p} [f_m(X_i)]_+ \tag{4.9}$$

denote the distance between w and H_0 over the set S_ϑ. For the next theorem, I will use ρ_ϑ-metric (instead of ρ-metric) to measure the distance between alternatives and the null hypothesis. Such restrictions are quite common in the literature. See, for example, Dumbgen and Spokoiny (2001) and Lee et al. (2011). Let $\mathcal{G}_\vartheta$ be the subset of all elements of $\mathcal{G}$ such that $\inf_{w \in \mathcal{G}_\vartheta} \rho_\vartheta(w, H_0) \geq C h_{\min}^\zeta$ for some sufficiently large constant C if $\zeta \leq d$ and $\inf_{w \in \mathcal{G}_\vartheta} \rho_\vartheta(w, H_0)(n/\log n)^{\zeta/(2\zeta+d)} \to \infty$ if $\zeta > d$. Then

Theorem 4. *Let assumptions 4-8 hold. For $P = PIA$ or RMS, if (i) $\zeta \leq d$ or (ii) $\zeta > d$ and $h_{\min} < (\log n/n)^{1/(2\zeta+d)}$ for sufficiently large n, then*

$$\sup_{w \in \mathcal{G}_\vartheta} P_w\{\hat{T} \leq c_{1-\alpha}^P\} \to 0 \tag{4.10}$$

as $n \to \infty$.

Remark 3. Recall the CMI model from the second example mentioned in the introduction where $m(X, W, \theta) = \tilde{m}(X, W) + \theta$. Assume that $X \in \mathbb{R}$ and $E[\tilde{m}(X, W)|X] = -|X|^\nu$ with $\nu > 1$. In this model, the identified set is $\Theta_I = \{\theta \in \mathbb{R} : \theta \leq 0\}$. The theorem above shows that the test developed in this paper is consistent against sequences of alternatives $\theta_0 = \theta_{0,n}$ whenever $(n/\log n)^{\nu/(2\nu+1)}\theta_{0,n} \to \infty$. At the same time, it follows from Armstrong (2011a) that the test of Andrews and Shi (2010) is consistent only if $n^{\nu/(2(\nu+1))}\theta_{n,0} \to \infty$, so their test has a slower rate of consistency than that developed in this paper by a polynomial order.

4.6 Lower Bound on the Minimax Rate of Testing

In this section, I give a lower bound on the minimax rate of testing. For any $X = \{X_i\}_{i=1}^\infty$ satisfying assumption 1, let $\mathcal{G}_X$ denote the set of all models $w = (f, \{\varepsilon_i\}_{i=1}^\infty, X\}$ in $\mathcal{G}$. For given X and S_ϑ defined in the previous section, let $N(h, S_\vartheta)$ be the largest m such that there

exists $\{x_1, ..., x_m\} \subset S_\vartheta$ with $\|x_i - x_j\| \geq h$ for all $i, j = 1, ..., m$ if $i \neq j$. I will assume that $N(h, S_\vartheta) \geq Ch^{-d}$ for all $h \in (0, 1)$ and sufficiently large n for some constant $C > 0$. In an iid setting, this condition holds a.s. under the conditions of lemma 4. Let $\phi_n(Y_1, ..., Y_n)$ denote a sequence of tests, where $\phi_n(Y_1, ..., Y_n)$ denotes the probability of rejecting the null hypothesis upon observing sample $Y = (Y_1, ..., Y_n)$.

Theorem 5. *Let assumptions 4-8 hold. Assume that (i) $N(h, S_\vartheta) \geq Ch^{-d}$ for all $h \in (0, 1)$ and sufficiently large n for some constant $C > 0$, (ii) $\varsigma = [\tau]$, and (iii) $r_n(n/\log n)^{\tau/(2\tau+d)} \to 0$ as $n \to \infty$ for some sequence of positive numbers r_n. Then for any sequence of tests $\phi_n(Y_1, ..., Y_n)$ with $\sup_{w \in \mathcal{G}_0 \cap \mathcal{G}_X} E_w[\phi_n(Y_1, ..., Y_n)] \leq \alpha$,*

$$\limsup_{n \to \infty} \inf_{w \in \mathcal{G}_X, \rho_\vartheta(w, H_0) \geq Cr_n} E_w[\phi_n(Y_1, ..., Y_n)] \leq \alpha \tag{4.11}$$

Remark 4. Since $\mathcal{F}_{[\tau]}(\tau, L) \subset \mathcal{F}(\tau, L)$, the same lower bound applies for the class $\mathcal{F}(\tau, L)$ as well. The same lower bound also applies with $\mathcal{G}$ instead of $\mathcal{G}_X$. Comparing this result with theorem 4 shows that the test presented in this paper is minimax rate optimal (for almost all sequences $\{X_i\}_{i=1}^\infty$) if $\zeta = \tau > d$ and $h_{\min}$ is chosen to converge to zero fast enough. When $\zeta = \tau = d$ and β_n is set to be constant, the test is minimax rate optimal up to some logarithmic factors if $h_{\min}$ is chosen to converge to zero as fast as possible satisfying assumption 4(ii). When $\tau < d$, the test is not minimax rate optimal since the rate of consistency does not match the lower bound[8].

5 Extentions

In this section, I briefly outline two extentions of the test developed in this paper. One of them concerns with the case of infinitely many CMI. The other one deals with local CMI. For brevity, I only discuss basic results. In both cases, I am interested in testing the null hypothesis, H_0, that $\theta = \theta_0$ against the alternative, H_a, that $\theta \neq \theta_0$.

Infinitely Many CMI. In many cases the parameter θ is restricted by a countably infinite number of CMI, i.e. $p = \infty$. For example, recall the English auction model and the model with interval data from section 2. In those models, inequalities (2.1) and (2.3)-(2.4) hold for all $v \in \mathbb{R}$. Taking rational values of v leads to a countably infinite number of CMI. Note that the last step does not change the identified set if left-hand sides of these inequalities are continuous in v or, at least, right or left continuous.

Let $\tilde{m} : \mathbb{R}^d \times \mathbb{R}^k \times \Theta \to \mathbb{R}^{\mathbb{N}}$ be some known function where $\mathbb{N}$ denotes the set of natural

[8]It is easy to show that the lower bound is achieved by the test with a higher order kernel, so the lower bound is tight.

numbers. Suppose that $\theta \in \Theta$ satisfies

$$E[\tilde{m}(X, W, \theta)|X] \leq 0 \, a.s. \tag{5.1}$$

Given θ_0, define $\tilde{f}(X) = E[\tilde{m}(X, W, \theta_0)|X]$. In addition, denote $\tilde{\varepsilon}_i = \tilde{m}(X_i, W_i, \theta_0) - \tilde{f}(X_i)$, and $\tilde{\Sigma}_i = E[\tilde{\varepsilon}_i \tilde{\varepsilon}_i^T | X_i]$. Let $\mathcal{G}$ denote the set of all models $(\tilde{f}, \{\tilde{\varepsilon}_i\}_{i=1}^{\infty}, \{X_i\}_{i=1}^{\infty})$ satisfying assumptions 1-3 with $\tilde{\varepsilon}$, $\tilde{\Sigma}$, and $\tilde{f}$ instead of ε, Σ, and f correspondingly and $p = \infty$. Let $\mathcal{G}_0$ denote the set of models $w = (\tilde{f}, \{\tilde{\varepsilon}_i\}_{i=1}^{\infty}, \{X_i\}_{i=1}^{\infty})$ in $\mathcal{G}$ satisfying $\tilde{f} \leq 0$ a.s., $\mathcal{G}_{00}$ denote the set of models w in $\mathcal{G}_0$ satisfying $\tilde{f} = 0$ a.s., and $\mathcal{G}_\rho$ denote the set of models w in $\mathcal{G}$ satisfying $\rho(w, H_0) \geq \rho$ where $\rho(w, H_0)$ is defined as in (4.6) with $\tilde{f}$ instead of f and $p = \infty$. Consider the test based on the first $Q = Q_n$ inequalities. More precisely, let $m : \mathbb{R}^d \times \mathbb{R}^k \times \Theta \to \mathbb{R}^Q$ be the vector-valued function whose j-th component coincides with j-th component of $\tilde{m}$ for all $j = 1, ..., Q$, and consider the test described in section 3 based on inequalities $E[m(X, W, \theta)|X] \leq 0$ a.s. Denote its critical value by $c_{1-\alpha}^P$ with $P = PIA$ or RMS. It will be assumed that $Q_n \to \infty$ as $n \to \infty$. An advantage of the finite sample approach used in this paper is that it immediately gives certain conditions that ensure that such a test maintains the required size as $n \to \infty$.

Corollary 1. *Let assumptions 4(i), 5, 7, and 8 hold with Q_n instead of p. In addition, assume that (i) $\max_{i=1,...,n} \|\hat{\Sigma}_i - \Sigma_i\|_o = o_p(n^{-\kappa})$ for some $\kappa > 0$[9], (ii) $Q_n \to \infty$, (iii) $(Q_n \log n)^{1/2}/n^{\kappa/4} \to 0$ as $n \to \infty$, and (iv) $n^{1-\epsilon} h_{\min}^{3d} > Q_n^6 C_5$ for all n. Then for $P = PIA$ or RMS,*

$$\inf_{w \in \mathcal{G}_0} P_w\{\hat{T} \leq c_{1-\alpha}^P\} \geq 1 - \alpha + o(1) \tag{5.2}$$

as $n \to \infty$. In addition,

$$\sup_{w \in \mathcal{G}_{00}} P_w\{\hat{T} \leq c_{1-\alpha}^P\} = 1 - \alpha + o(1) \tag{5.3}$$

Finally,

$$P_w\{\hat{T} \leq c_{1-\alpha}^P\} \to 0 \tag{5.4}$$

for any $w \in \mathcal{G}_\rho$ with $\rho > 0$.

Remark 5. (i) This corollary shows that the randomized test has the correct asymptotic size, is asymptotically not conservative, and is consistent against fixed alternatives outside of the set Θ_I.

(ii) Note that κ appearing in conditions (i) and (iii) in this corollary will generally be different from κ used in assumption 6 because of increasing number of moment functions.

[9] All quantities undefined in this section coincide symbol-by-symbol with those used in sections 3 and 4.

Local CMI. Suppose that the parameter θ is restricted by the following inequalities:

$$E[m(X, W, \theta)|X,\ Z = z_0] \leq 0\, a.s. \tag{5.5}$$

where $m(\cdot,\cdot,\cdot)$, X, and W are as above, Z is some random $\mathbb{R}^{d_z}$-dimensional random vector, which may or may not include some components of W, and z_0 is some fixed point. CMI of the form (5.5) arise in nonparametric and semiparametric inference. For example, recall the English auction model from section 2. In that model, suppose that the set of covariates is (X, Z) instead of X so that $F = F(v, X, Z)$. Suppose that the case $Z = z_0$ is of interest. Denote $\tilde{F}(v, X) = F(v, X, z_0)$. Then inequality (2.1) leads to

$$E[\phi^{-1}(\tilde{F}(v, X)) - I\{b_{i:m} \leq v\}|X,\ Z = z_0] \leq 0\, a.s. \tag{5.6}$$

Parameterizing the function $\tilde{F}(\cdot,\cdot)$ gives inequalities of the form (5.5). Note that parameterizing $\tilde{F}(\cdot,\cdot)$ instead of $F(\cdot,\cdot,\cdot)$ reduces the risk of misspecification, which makes this approach attractive when the only interesting value of Z is z_0.

Given θ_0, define $f^{x,z}(X, Z) = E[m(X, W, \theta_0)|X, Z]$. In addition, denote $\varepsilon_i^{x,z} = m(X_i, W_i, \theta_0) - f^{x,z}(X_i, Z_i)$, and $\Sigma_i^{x,z} = E[\varepsilon_i^{x,z}(\varepsilon_i^{x,z})^T|X_i, Z_i]$. Let $\hat{\Sigma}_i^{x,z}$ be an estimator of $\Sigma_i^{x,z}$ $(i = 1, ..., n)$ as described in section 3.3. Let N be a subset of all observations $i = 1, ..., n$ such that $\|Z_i - z_0\| < a$ for all $i \in N$. It will be assumed that $a = a_n \to 0$ as $n \to \infty$. Denote the number of elements in N by n_a. Without loss of generality, I assume that observations in N are those corresponding to $i = 1, ..., n_a$. Let $\mathcal{G}$ denote the set of models $(f^{x,z}, \{\varepsilon_i^{x,z}\}_{i=1}^{\infty}, \{(X_i, Z_i)\}_{i=1}^{\infty})$ satisfying assumptions 1-3 with $\varepsilon_i^{x,z}$, $\Sigma_i^{x,z}$, $f^{x,z}$, $d + d_z$, and (X_i, Z_i) instead of ε_i, Σ_i, f, d, and X_i correspondingly. Let $\mathcal{G}_0$ denote the set of all models w in $\mathcal{G}$ satisfying $f^{x,z}(X, z_0) \leq 0$ a.s., $\mathcal{G}_{00}$ denote the set of models w in $\mathcal{G}_0$ satisfying $f^{x,z}(X, z_0) = 0$ a.s. Denote $\mathcal{N}_a = \{(i, m) :\ i = 1, ..., \infty,\ m = 1, ..., p,\ \|Z_i - z_0\| \leq a\}$. Define the distance between the model $w = (f^{x,z}, \{\varepsilon_i^{x,z}\}_{i=1}^{\infty}, \{(X_i, Z_i)\}_{i=1}^{\infty}) \in \mathcal{G}$ and the null hypothesis by

$$\rho_z(w, H_0) = \inf_{a \in (0,\infty)} \sup_{(i,m) \in \mathcal{N}_a} [f_m^{x,z}(X_i, Z_i)]_+ \tag{5.7}$$

Let $\mathcal{G}_{z,\rho}$ denote the set of all models w in $\mathcal{G}$ satisfying $\rho_z(w, H_0) \geq \rho > 0$.

Consider the test described in section 3 based on the data $\{(X_i, W_i)\}_{i \in N}$ with $\hat{\Sigma}_i^{x,z}$ instead of $\hat{\Sigma}_i$ $(i = 1, ..., n_a)$ as if we would like to test the null hypothesis that $\theta = \theta_0$ in the model $E[m(X, W, \theta)|X] \leq 0$ with n_a observations. Note that this test uses Z_i only for selecting N and estimating $\hat{\Sigma}_i^{x,z}$. Denote its critical value by $c_{1-\alpha}^P$ with $P = PIA$ or RMS.

Corollary 2. *Let assumption 6 hold with $\hat{\Sigma}_i^{x,z}$ and $\Sigma_i^{x,z}$ instead of $\hat{\Sigma}_i$ and Σ_i correspondingly. In addition, let assumptions 4, 5, 7, and 8 hold with n_a instead of n. Finally, assume that (i)*

$a = a_n \to 0$, (ii) for some constant $C > 0$, $n_a \geq Cna_n^{d_z}$, (iii) for some constants $0 < C_1 < C_2 < \infty$, the number of elements in the set $\{X_j : \|X_j - X_i\| \leq h, j \in N\}$ is bounded away from zero and from above by $C_1 n_a h^d$ and $C_2 n_a h^d$ correspondingly for all $i \in \mathbb{N}$ and $h \in H_n$. Then for $P = PIA$ or RMS,

$$\inf_{w \in \mathcal{G}_0} P_w\{\hat{T} \leq c_{1-\alpha}^P\} \geq 1 - \alpha + o(1) \tag{5.8}$$

as $n \to \infty$ provided that in addition to (i)-(iii) we have (iv) $na_n^{d_z+2} h_{\max}^d \log n \to 0$, and (v) for some constant $C > 0$, $f_m^{x,z}(X_i, Z_i) \leq Ca_n$ for all $i \in N$ and $m = 1, ..., p$. Further,

$$\sup_{w \in \mathcal{G}_{00}} P_w\{\hat{T} \leq c_{1-\alpha}^P\} \geq 1 - \alpha + o(1) \tag{5.9}$$

as $n \to \infty$ provided that in addition to (i)-(v) we have (vi) for some constant $C > 0$, $f_m^{x,z}(X_i, Z_i) \geq -Ca_n$ for all $i \in N$ and $m = 1, ..., p$. Finally,

$$P_w\{\hat{T} \leq c_{1-\alpha}^P\} \to 0 \tag{5.10}$$

for any $w \in \mathcal{G}_{z,\rho}$ with $\rho > 0$.

Remark 6. (i) Note that in an iid setting, if $f^{x,z}(x, z_0) > 0$ for some x such that (x, z_0) is inside of the support of (X, Z), then it follows as in the proof of lemma 4 that $\rho_z(w, H_0) > 0$ a.s. So, the corollary above shows that the test has correct asymptotic size, is asymptotically not conservative, and is consistent against any fixed alternative outside of the set Θ_I.

(ii) Note that the corollary remains valid if $h_{\max} \to 0$ as $n \to 0$.

(iii) Condition (iv) in this corollary requires that $na^{d_z+2} h_{\max}^d \log n \to 0$. This condition ensures that the bias due to using data with $Z_i \neq z_0$ is asymptotically negligible. Given that small values of a lead to small effective sample size n_a while small values of $h_{\max}$ lead to large variance of the kernel estimator, it is useful to set $h_{\max} \to 0$ as $n \to \infty$ to balance these effects.

6 Monte Carlo Results

In this section, I present results of two Monte Carlo simulation studies. The aim of these simulations is twofold. First, I demonstrate that my test accurately maintains size in finite samples. Second, I compare relative advantages and disadvantages of my test and the tests of Andrews and Shi (2010), Chernozhukov et al. (2009), and Lee et al. (2011). The methods of Andrews and Shi (2010) and Lee et al. (2011) are most appropriate for detecting flat alternatives, which represent one-dimensional local alternatives. These methods have low power

against alternatives with peaks, however. The test of Chernozhukov et al. (2009) has higher power against such alternatives, but it requires knowing smoothness properties of the moment functions. The authors suggest certain rule-of-thumb techniques to choose a bandwidth value. Finally, the main advantage of my test is its adaptiveness. In comparison with Andrews and Shi (2010) and Lee et al. (2011), my test has higher power against alternatives with peaks. In comparison with Chernozhukov et al. (2009), my test has higher power when their rule-of-thumb techniques lead to an inappropriate bandwidth value. For example, this happens when the underlying moment function is mostly flat but varies significantly in the region where the null hypothesis is violated (the case of spatially inhomogeneous alternatives, see Lepski and Spokoiny (1999)).

First simulation study. The data generating process is

$$Y = L(M - |X|)_+ - m + \varepsilon \tag{6.1}$$

where X, Y, and ε are scalar random variables and L, M, and m are some constants. X is distributed uniformly on $(-2, 2)$. Depending on the experiment, ε is distributed according to $0.1 \cdot N(0,1)$ or $(\xi \cdot 0.07 + (1-\xi) \cdot 0.18) \cdot N(0,1)$ where ξ is a Bernoulli random variable with $p(\xi = 1) = 0.8$ and $p(\xi = 0) = 0.2$ and is independent of $N(0,1)$. In both cases, ε is independent of X. I consider the following specifications for parameters. Case 1: $L = M = m = 0$. Case 2: $L = 0.1$, $M = 0.2$, $m = 0.02$. Case 3: $L = M = 0$, $m = -0.02$. Case 4: $L = 2$, $M = 0.2$, $m = 0.2$. Note that $E[Y|X] \leq 0$ a.s. in cases 1 and 2 while $P\{E[Y|X] > 0\} > 0$ in cases 3 and 4. In case 3, the alternative is flat. In case 4, the alternative has a peak in the region where the null hypothesis is violated. I have chosen parameters so that rejection probabilities are strictly greater than 0 and strictly smaller than 1 in most cases so that meaningful comparisons are possible. I generate samples $(X_i, Y_i)_{i=1}^n$ of size $n = 250$ and 500 from the distribution of (X, Y). In all cases, I consider tests with the nominal size 10%. The results are based on 1000 simulations for each specification.

For the test of Andrews and Shi (2010), I consider their Kolmogorov-Smirnov test statistic with boxes and truncation parameter 0.05. I simulate both plugin (AS, plugin) and GMS (AS, GMS) critical values based on the bootstrap suggested in their paper. I use the support of the empirical distribution of X to choose a set of weighting functions. All other tuning parameters are set as prescribed in their paper.

Implementing all other tests requires selecting a kernel function. In all cases, I use

$$K(x) = 1.5(1 - 4x^2)_+ \tag{6.2}$$

For the test of Chernozhukov et al. (2009), I use their kernel type test statistic with critical

values based on the multiplier bootstrap both with (CLR, $\hat{V}$) and without (CLR, V) the set estimation. Both Chernozhukov et al. (2009) and Lee et al. (2011) (LSW) circumvent edge effects of kernel estimators by restricting their test statistics to the proper subsets of the support of X. To accomodate this, I select the 10%th and 90%th percentiles of the empirical distribution of X as bounds for the set over which the test statistics are calculated. Both tests are nonadaptive. In particular, there is no formal theory on how to choose bandwidth values in their tests, so I follow their informal suggestions. For the test of Lee et al. (2011), I use their test statistic based on one-sided L_1-norm.

Let me now describe the choice of parameters for the test developed in this paper. The largest bandwidth value, $h_{\max}$, is set to be one half of the length of the support of the empirical distribution. I choose the smallest bandwidth value, $h_{\min}$, so that the kernel estimator uses on average 15 data points when $n = 250$ and 20 data points when $n = 500$, which roughly corresponds to my recommendations in section 3. The scaling parameter, a, equals 0.8 so that the set of bandwidth values is

$$H_n = \{h = h_{\max} 0.8^k : h \geq h_{\min}, k = 0, 1, 2, ...\} \tag{6.3}$$

My test requires selecting the set S_n. For each bandwidth value, h, I select the largest subset, $S_{n,h}$, of X_i's such that $X_i - X_j \geq h$ for any nonequal elements in $S_{n,h}$, and the smallest X_i is always in $S_{n,h}$. Then $S_n = \{(i, h) : h \in H_n, X_i \in S_{n,h}\}$. I estimate Σ_i using the method of Rice (1984). Specifically, I rearrange the data so that $X_1 \leq ... \leq X_n$ and set $\hat{\Sigma}_i = \hat{\Sigma} = \sum_{i=2}^{n} (Y_i - Y_{i-1})^2/(2n)$. Finally, for the RMS critical value, I set $\gamma = 0.1/\log(n)$ to make meaningful comparisons with the test of Chernozhukov et al. (2009). In all bootstrap procedures, for all tests, I use 1000 repetitions when $n = 250$ and 500 repetitions when $n = 500$.

The results of the first simulation study are presented in table 1 for $n = 250$ and in table 2 for $n = 500$. In both tables, my test is denoted as Adaptive test with plug-in and RMS critical values. Consider first results for $n = 250$. In case 1, where the null hypothesis holds, all tests have rejection probabilities close to the nominal size 10% both for normal and mixture of normals disturbances. In case 2, where the null hypothesis holds but the underlying regression function is mainly strictly below the borderline, all tests are conservative. When the null hypothesis is violated with a flat alternative (case 3), the tests of Andrews and Shi (2010) and Lee et al. (2011) have highest rejection probabilities as expected from the theory. In this case, my test is less powerful in comparison with these tests and somewhat similar to the method of Chernozhukov et al. (2009). This is compensated in case 4 where the null hypothesis is violated with the peak-shaped alternative. In this case, the power of my test is

Table 1: Results of Monte Carlo Experiments, $n = 250$

Distribution ε	Case	Probability of Rejecting Null Hypothesis						
		AS, plugin	AS, GMS	LSW	CLR, V	CLR, $\hat{V}$	Adaptive test, plugin	Adaptive test, RMS
Normal	1	0.099	0.102	0.124	0.151	0.151	0.101	0.101
	2	0.002	0.007	0.000	0.008	0.008	0.009	0.009
	3	0.910	0.910	0.941	0.808	0.808	0.723	0.723
	4	0.000	0.143	0.000	0.122	0.191	0.589	0.821
Mixture	1	0.078	0.086	0.107	0.134	0.134	0.124	0.124
	2	0.002	0.002	0.000	0.010	0.010	0.016	0.016
	3	0.904	0.905	0.925	0.833	0.833	0.692	0.692
	4	0.000	0.121	0.000	0.111	0.197	0.555	0.808

Table 2: Results of Monte Carlo Experiments, $n = 500$

Distribution ε	Case	Probability of Rejecting Null Hypothesis						
		AS, plugin	AS, GMS	LSW	CLR, V	CLR, $\hat{V}$	Adaptive test, plugin	Adaptive test, RMS
Normal	1	0.095	0.104	0.119	0.126	0.126	0.103	0.103
	2	0.000	0.001	0.000	0.002	0.002	0.008	0.008
	3	0.997	0.997	0.996	0.954	0.954	.903	0.903
	4	0.008	0.587	0.000	0.497	0.694	0.976	0.999
Mixture	1	0.120	0.123	0.130	0.117	0.117	0.119	0.119
	2	0.000	0.001	0.000	0.000	0.000	0.010	0.010
	3	0.993	0.993	0.996	0.949	0.949	0.903	0.903
	4	0.005	0.549	0.000	0.456	0.625	0.978	0.997

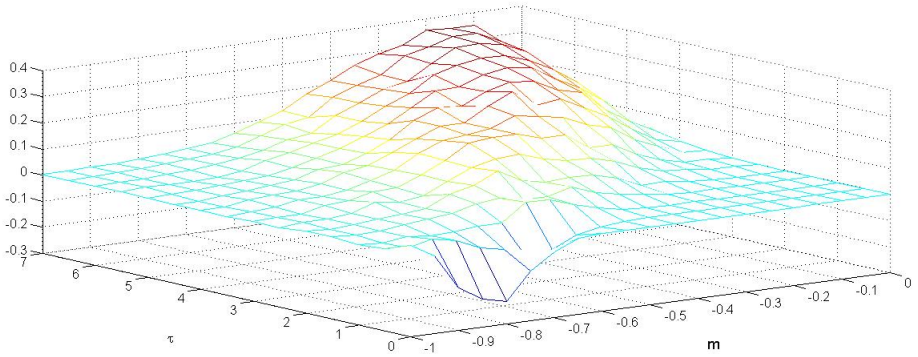
much higher than that of competing tests. This is especially true for my test with RMS critical values whose rejection probability exceeds 80% while rejection probabilities of competing tests do not exceed 20%. Note that all results are stable across distributions of disturbances. Also note that my test with RMS critical values has much higher power than the test with plugin critical values in case 4. So, among these two tests, I recommend the test with RMS critical values. Results for $n = 500$ indicate a similar pattern.

Second simulation study. In the second simulation study, I compare the power function of the test developed in this paper with that of the Andrews and Shi's (2010) test, which is most closely related to my method. For my test, I use the RMS critical value. For the test of Andrews and Shi (2010), I use their GMS critical value. The data generating process is

$$Y = m + \sqrt{2\pi}\phi(\tau X) + \varepsilon \tag{6.4}$$

where X, Y, and ε are scalar random variables, m and τ are some constants, and $\phi(\cdot)$ is the pdf of the standard Gaussian distribution. X is distributed uniformly on $(-2, 2)$, and ε

Figure 1: The difference between the rejection probabilities of the test developed in this paper and of the test of Andrews and Shi (2010) (with RMS and GMS critical values correspondingly). The nominal size is 10%. Results are based on 500 simulations. The figure shows that the rejection probability of the test developed in this paper is higher than that of the test of Andrews and Shi (2010) in most cases and is strictly higher over a wide region of parameter values.



has $N(0,1)$ distrubution. In this simulation study, I use samples $(X_i, Y_i)_{i=1}^n$ of size $n = 250$ from the distribution of (X, Y). Both tests are based on the same specifications as in the first simulation study except that now I use 100 repetitions for all bootstrap procedures in order to conserve computing time. At each point, the rejection probabilities are estimated using 500 simulations.

Note that τ is naturally bounded from below because τ and $-\tau$ yield the same results. So, I set $\tau \geq 0$. In addition, $E[Y|X] \leq 0$ a.s. if $m \leq -1$. Therefore, I set $m \geq -1$. Figure 1 shows the difference between the rejection probabilities of my test and of the test of Andrews and Shi (2010). This figure shows that the rejection probability of the test developed in this paper is higher than that of the test of Andrews and Shi (2010) in most cases and is strictly higher over a wide region of parameter values. The exception is a narrow region where τ is close to 0 (flat alternatives) and m is close to -1. Concluding this section, I note that all simulation results are consistent with the presented theory.

7 Conclusions

In this paper, I develop a new test of conditional moment inequalities. In contrast to some other tests in the literature, my test is directed against general nonparametric alternatives yielding high power in a large class of CMI models. Considering kernel estimates of moment functions with many different values of the bandwidth parameter allows me to construct a test that automatically adapts to the unknown smoothness of moment functions and selects the most appropriate testing bandwidth value. The test developed in this paper has uniformly

correct asymptotic size, no matter whether the model is identified, weakly identified, or not identified, is consistent against any fixed alternative outside of the set Θ_I, and is uniformly consistent against certain, but not all, large classes of smooth alternatives whose distance from the null hypothesis converges to zero at a fastest possible rate. The tests of Andrews and Shi (2010) and Lee et al. (2011) have nontrivial power against $n^{-1/2}$-local one-dimensional alternatives whereas my method only allows for nontrivial testing against $(n/\log n)^{-1/2}$-local alternatives of this type. The additional $(\log n)^{1/2}$ factor should be regarded as the price for having fast rate of uniform consistency. There exist sequences of local alternatives against which their tests are not consistent whereas mine is. Monte Carlo experiments give an example of a CMI model where finite sample power of my test greatly exceeds that of competing tests.

A Appendix

This Appendix contains proofs of all results stated in the main part of the paper. Section A.1 gives a proof of the uniform consistency of the estimator $\hat{\Sigma}_i$ of Σ_i described in section 3.3. I provide the proof because I was not able to find it in the literature. Section A.2 derives a bound on the modulus of continuity in the spectral norm of the square root operator on the space of symmetric positive semidefinite matrices. Section A.3 gives a straighforward generalization of results in Chatterjee (2005) on stochastic approximation to the case of multidimensional random variables. Those results have their own value as they can be used as an alternative to results in empirical process theory. They are also useful because they give an explicit bound on the approximation error. Section A.4 gives sufficient conditions for assumption 1 in the main part of the paper. Section A.5 presents an anticoncentration inequality for the maximum of Gaussian random variables with unit variance. Section A.6 describes a result on Gaussian random variables that is used in the proof of the lower bound on the minimax rate. Section A.7 develops some preliminary technical results necessary for the proofs of the main theorems. Finally, section A.8 presents the proofs of the theorems stated in the main part of the paper.

In this appendix, I use C and its variants to denote generic constants that are independent of n.

A.1 Lemma on the Estimator of Σ_i

Lemma 1. *Let $\hat{\Sigma}_i$ be an estimator of Σ_i described in section 3.3. Let assumptions 1-3 hold. In addition, assume that (i) $E[|\varepsilon_j|^{4+\delta}] < C$ for all $j = 1, ..., n$ and some $C > 0$, (ii) $b \leq n^{-C}$ for some $C > 0$, (iii) $\min_{i=1,...,n} |J(i)|/n^{1/(2+\delta)} \geq n^C$ for some $C > 0$, (iv) $\|\Sigma_i - \Sigma_j\|_o \leq C\|X_i - X_j\|$ for some $C > 0$. Then there exists some $\kappa > 0$ such that $\max_{i=1,...,n} \|\hat{\Sigma}_i - \Sigma_i\|_o = o_p(n^{-\kappa})$.*

Remark 7. Note that under assumptions of lemma 4, condition (iii) above follows from $n^{(1+\delta)/(2+\delta)}b^d \geq n^C$, which is an elementary condition.

Proof. By definition,

$$\hat{\Sigma}_i = \sum_{j \in J(i)} (Y_{k(j)} - Y_j)(Y_{k(j)} - Y_j)^T / (2|J(i)|) \tag{A.1}$$

Since all norms on the finite-dimensional linear space are equivalent (theorem 1.6 in Kress (1999)), it is enough to prove that

$$\max_{i=1,...,n} |\hat{\Sigma}_{i,m_1m_2} - \Sigma_{i,m_1m_2}| = o_p(n^{-\kappa}) \tag{A.2}$$

for all $m_1, m_2 = 1, ..., p$. The proof will be given for $m_1 = m_2 = 1$. The result for all other m_1, m_2 follows from the same argument. To simplify notation, I will write Σ_i, $\hat{\Sigma}_i$, $f(X_i)$, and ε_i instead of $\Sigma_{i,11}$, $\hat{\Sigma}_{i,11}$, $f_1(X_i)$, and $\varepsilon_{i,1}$ correspondingly as if it were a one-dimensional case.

Denote $\tilde{\varepsilon}_i = \varepsilon_i I\{\varepsilon_i \leq M\}$ for $M = n^{1/(4+\delta/2)}$. Since $E[|\varepsilon_i|^{4+\delta}] < C$, it follows that $E[\max_{i=1,...,n} |\varepsilon_i|] \leq Cn^{1/(4+\delta)}$ for some (possibly different) $C > 0$ (see lemma 2.2.2 in Van der Vaart and Wellner (1996)). Then Markov inequality gives

$$P\{\max_{i=1,...,n} |\varepsilon_i| > M\} \leq Cn^{1/(4+\delta)}/M \to 0 \tag{A.3}$$

So,

$$P_1 = P\{\max_{i=1,...,n} |\tilde{\varepsilon}_i - \varepsilon_i| > 0\} \to 0 \tag{A.4}$$

Denote $\tilde{\Sigma}_i = E[\tilde{\varepsilon}_i^2]$ $(i = 1, ..., n)$. Then $\tilde{\Sigma}_i = \Sigma_i - E[\varepsilon_i^2 I\{\varepsilon_i > M\}]$. Combining Fubini theorem and Markov inequality yields

$$\begin{aligned} E[\varepsilon_i^2 I\{\varepsilon_i > M\}] &= \int_0^\infty P\{\varepsilon_i^2 I\{\varepsilon_i > M\} > t\} dt \\ &\leq MP\{\varepsilon_i > M\} + \int_M^\infty E[\varepsilon_i^4]/t^2 dt \\ &\leq E[\varepsilon_i^4](1/M^3 + 1/M) \\ &\leq 2E[\varepsilon_i^4]/M \end{aligned}$$

In addition, denote $\tilde{Y}_i = f(X_i) + \tilde{\varepsilon}_i$ and

$$\bar{\Sigma}_i = \sum_{j \in J(i)} (\tilde{Y}_{k(j)} - \tilde{Y}_j)(\tilde{Y}_{k(j)} - \tilde{Y}_j)^T / (2|J(i)|) \tag{A.5}$$

$(i = 1, ..., n)$. Then

$$P\{\max_{i=1,...,n} |\bar{\Sigma}_i - \hat{\Sigma}_i| > 0\} = P_1 \to 0 \tag{A.6}$$

Note that for sufficiently small κ ($\kappa < 1/(4 + \delta/2)$),

$$P\{\max_{i=1,...,n} |\hat{\Sigma}_i - \Sigma_i| > n^{-\kappa}\} \leq P\{\max_{i=1,...,n} |\bar{\Sigma}_i - \tilde{\Sigma}_i| > n^{-\kappa}/2\} + o(1) \tag{A.7}$$

as $n \to \infty$. By the union bound,

$$P\{\max_{i=1,...,n} |\bar{\Sigma}_i - \tilde{\Sigma}_i| > n^{-\kappa}/2\} \leq \sum_{i=1}^{n} P\{|\bar{\Sigma}_i - \tilde{\Sigma}_i| > n^{-\kappa}/2\} \tag{A.8}$$

Then

$$P\{|\bar{\Sigma}_i - \tilde{\Sigma}_i| > n^{-\kappa}/2\} \leq P_1 + P_2 + P_3 \tag{A.9}$$

where

$$P_1 = P\{\sum_{j \in J(i)} (f(X_{k(j)}) - f(X_j))^2/(2|J(i)|) > n^{-\kappa}/6\} \tag{A.10}$$

$$P_2 = P\{|\sum_{j \in J(i)} (f(X_{k(j)}) - f(X_j)(\tilde{\varepsilon}_{k(j)} - \tilde{\varepsilon}_j))|/|J(i)| > n^{-\kappa}/6\} \tag{A.11}$$

$$P_3 = \{|\sum_{j \in J(i)} (\tilde{\varepsilon}_{k(j)} - \tilde{\varepsilon}_j)^2/(2|J(i)|) - \tilde{\Sigma}_i| > n^{-\kappa}/6\} \tag{A.12}$$

By assumption 3, $|f(X_{k(j)}) - f(X_j)| \leq L\|X_{k(j)} - X_j\| \leq 2Lb$. Since b converges to zero at a polynomial rate, $P_1 = 0$ for sufficiently large n if κ is sufficiently small. Consider P_3. Note that $P_3 \leq P_{31} + P_{32}$ where

$$P_{31} = \{|\sum_{j \in J(i)} \tilde{\varepsilon}_j^2/|J(i)| - \tilde{\Sigma}_i| > n^{-\kappa}/12\} \tag{A.13}$$

and

$$P_{32} = \{|\sum_{j \in J(i)} \tilde{\varepsilon}_{k(j)}\tilde{\varepsilon}_j|/|J(i)| > n^{-\kappa}/12\} \tag{A.14}$$

Since $|\Sigma_i - \Sigma_j| \leq L\|X_i - X_j\|$, it follows that

$$P_{31} = \{|\sum_{j \in J(i)} (\tilde{\varepsilon}_j^2 - \tilde{\Sigma}_j)|/|J(i)| > n^{-\kappa}/24\} + o(1) \tag{A.15}$$

if κ is sufficiently small. Then Hoeffding inequality gives (see proposition 1.3.5 in Dudley (1999))

$$P_{31} \leq 2\exp\{-C|J(i)|/(M^2 n^{2\kappa})\} \tag{A.16}$$

Therefore, $nP_{31} \to 0$ as $n \to 0$ if $\min_{i=1,...,n} |J(i)|/M^2 > n^C$ for some $C > 0$, which holds by assumption (iii), and κ is sufficiently small.

Now consider P_{32}. Denote $U(i) = \{j \in J(i) : j < k(j)\}$. Apply Hoeffding inequality conditional on $\{\tilde{\varepsilon}_j\}_{j \in U(i)}$. Since $|\tilde{\varepsilon}_j| \leq M$ for all $j = 1, ..., n$, $nP_{32} \to 0$ like $nP_{31} \to 0$. Similar argument shows that $nP_2 \to 0$ as well. The result follows. □

A.2 Continuity of the Square Root Operator on the Set of Positive Semidefinite Matrices

Lemma 2. *Let A and B be $p \times p$-dimensional symmetric positive semidefinite matrices. Then $\|A^{1/2} - B^{1/2}\|_o \leq p^{1/2}\|A - B\|_o^{1/2}$ where $\|\cdot\|_o$ means the spectral norm corresponding to the Euclidean norm on $\mathbb{R}^p$.*

Proof. Let $a_1, ..., a_p$ and $b_1, ..., b_n$ be orthogonal eigenvectors of matrices A and B correspondingly. Without loss of generality, I can and will assume that $\|a_i\| = \|b_i\| = 1$ for all $i = 1, ..., p$ where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^p$. Let $\lambda_1(A), ..., \lambda_p(A)$ and $\lambda_1(B), ..., \lambda_p(B)$ be corresponding eigenvalues. Let $f_{i1}, ..., f_{ip}$ be coordinates of a_i in the basis $(b_1, ..., b_p)$ for all $i = 1, ..., p$. Then $\sum_{j=1}^p f_{ij}^2 = 1$ for all $i = 1, ..., p$.

For any $i = 1, ..., p$,

$$\begin{aligned}
\sum_{j=1}^p (\lambda_i(A) - \lambda_j(B))^2 f_{ij}^2 &= \|\sum_{j=1}^p (\lambda_i(A) - \lambda_j(B)) f_{ij} b_j\|^2 \\
&= \|\lambda_i(A) a_i - \sum_{j=1}^p \lambda_j(B) f_{ij} b_j\|^2 \\
&= \|(A - B) a_i\|^2 \\
&\leq \|A - B\|_o^2
\end{aligned}$$

since $\|(A - B)a_i\| \leq \|A - B\|_o \|a_i\| = \|A - B\|_o$.

For $P = A, B$, $P^{1/2}$ has the same eigenvectors as P with corresponding eigenvalues equal

to $\lambda_1^{1/2}(P), ..., \lambda_n^{1/2}(P)$. Therefore, for any $i = 1, ..., p$,

$$\begin{aligned}
\|(A^{1/2} - B^{1/2})a_i\|^2 &= \sum_{j=1}^{p} (\lambda_i^{1/2}(A) - \lambda_j^{1/2}(B))^2 f_{ij}^2 \\
&\leq \sum_{j=1}^{p} |\lambda_i(A) - \lambda_j(B)| f_{ij}^2 \\
&\leq \left(\sum_{j=1}^{p} (\lambda_i(A) - \lambda_j(B))^2 f_{ij}^2 \right)^{1/2} \\
&\leq \|A - B\|_o
\end{aligned}$$

where the last line used the inequality derived above. For any $c \in \mathbb{R}^p$ with $\|c\| = 1$, let $d_1, ..., d_p$ be coordinates of c in the basis $(a_1, ..., a_p)$. Then

$$\begin{aligned}
\|(A^{1/2} - B^{1/2})c\| &= \|(A^{1/2} - B^{1/2}) \sum_{i=1}^{p} d_i a_i\| \\
&\leq \sum_{i=1}^{p} |d_i| \|(A^{1/2} - B^{1/2})a_i\| \\
&\leq \sum_{i=1}^{p} |d_i| \|A - B\|_o^{1/2} \\
&\leq p^{1/2} \|A - B\|_o^{1/2}
\end{aligned}$$

since $\sum_{i=1}^{p} d_i^2 = 1$. Thus, $\|A^{1/2} - B^{1/2}\|_o \leq p^{1/2} \|A - B\|_o^{1/2}$. □

A.3 Invariance Principle

In this section, I generalize results of Chatterjee (2005) to the case of random vectors ($p > 1$). I also specialize results for the case of linear functions because it allows to improve the rate derived in that paper. Let $Z_1, ..., Z_n$ be a sequence of independent p-dimensional random vectors with $E[Z_j] = 0$ for all $j = 1, ..., n$. Denote $Z = (Z_1, ..., Z_n)$. For all $k = 1, ..., K$ and $m = 1, ..., p$, let $f_{km}(Z) = \sum_{j=1}^{n} a_{kjm} Z_{j,m}$ be some linear function of Z where $a_{kjm} \geq 0$ for all $k = 1, ..., K$, $j = 1, ..., n$, and $m = 1, ..., p$, and $Z_{j,m}$ denotes m-th component of vector Z_j. Let $U_1, ..., U_n$ be a sequence of independent normal p-dimensional random vectors such that $E[U_j] = 0$ and $E[Z_j Z_j^T] = E[U_j U_j^T]$ for all $j = 1, ..., n$. Denote $U = (U_1, ..., U_n)$,

$$M = \max_{j,m} E[|Z_{j,m}|^3] + \max_{j,m} E[|U_{j,m}|^3] \tag{A.17}$$

and

$$C(g,\alpha) = \|g'''\|_\infty + 3\alpha\|g''\|_\infty + \alpha^2\|g'\|_\infty \tag{A.18}$$

for any thrice differentiable function $g : \mathbb{R} \to \mathbb{R}$ and $\alpha > 0$. Denote $a = \max_{k,j,m} a_{kjm}$. Then

Theorem 6. *For any thrice differentiable function g on $\mathbb{R}$ and $\alpha > 0$,*

$$\left| E\left[g(\max_{k,m} f_{km}(Z)) - g(\max_{k,m} f_{km}(U)) \right] \right| \leq 2\|g'\|_\infty \alpha^{-1}\log(Kp) + np^3a^3C(g,\alpha)M/6 \tag{A.19}$$

Remark. The constants in the inequality above can be improved somewhat by using expressions for A_1, A_2, and A_3 in the proof given below. I do not follow this step because that would mess up the statement of the theorem significantly.

Proof. As in Chatterjee (2005), for $\alpha \geq 1$, let $F_\alpha : \mathbb{R}^{p\times n} \to \mathbb{R}$ be such that

$$F_\alpha(x) = \alpha^{-1}\log(\sum_{k,m}\exp(\alpha f_{km}(x))) \tag{A.20}$$

for all $x \in \mathbb{R}^{p\times n}$. Then

$$\begin{aligned}
\max_{k,m} f_{km}(x) &= \alpha^{-1}\log(\exp(\alpha \max_{k,m} f_{km}(x))) \\
&\leq \alpha^{-1}\log(\sum_{k,m}\exp(\alpha f_{km}(x))) \\
&\leq \alpha^{-1}\log(Kp\exp(\alpha \max_{k,m} f_{km}(x))) \\
&\leq \alpha^{-1}\log(Kp) + \max_{k,m} f_{km}(x)
\end{aligned}$$

So,

$$|\max_{k,m} f_{km}(x) - F_\alpha(x)| \leq \alpha^{-1}\log(Kp) \tag{A.21}$$

Thus,

$$|E[g(\max_{k,m} f_{km}(Z))] - E[g(\max_{k,m} f_{km}(U))]| \leq$$
$$2\|g'\|_\infty\alpha^{-1}\log(Kp) + |E[g(F_\alpha(Z))] - E[g(F_\alpha(U))]|$$

For any $j = 0, ..., n$, denote $Z^j = (Z_1, ..., Z_j, U_{j+1}, ..., U_n)$ where $Z^0 = (U_1, ..., U_n)$ and $Z^n = (Z_1, ..., Z_n)$. Then

$$|E[g(F_\alpha(Z))] - E[g(F_\alpha(U))]| \leq \sum_{j=1}^{n} |E[g(F_\alpha(Z^j)] - E[g(F(Z^{j-1}))]| \tag{A.22}$$

For $Z_1, ..., Z_{j-1}, U_{j+1}, ..., U_n$ fixed, denote $l(Z_j) = g(F_\alpha(Z^j))$. By Taylor formula,

$$\begin{aligned} g(F_\alpha(Z^j)) - g(F_\alpha(Z^{j-1})) &= l(Z_j) - l(U_j) \\ &= \sum_{m_1} \frac{\partial l(0)}{\partial Z_{jm_1}}(Z_{jm_1} - U_{jm_1}) \\ &+ (1/2) \sum_{m_1,m_2} \frac{\partial^2 l(0)}{\partial Z_{jm_1}\partial Z_{jm_2}}(0)(Z_{jm_1}Z_{jm_2} - U_{jm_1}U_{jm_2}) \\ &+ (1/6) \sum_{m_1,m_2,m_3} \frac{\partial^3 l(\tilde{Z})}{\partial Z_{jm_1}\partial Z_{jm_2}\partial Z_{jm_3}} Z_{jm_1}Z_{jm_2}Z_{jm_3} \\ &- (1/6) \sum_{m_1,m_2,m_3} \frac{\partial^3 l(\tilde{U})}{\partial Z_{jm_1}\partial Z_{jm_2}\partial Z_{jm_3}} U_{jm_1}U_{jm_2}U_{jm_3} \end{aligned}$$

where $\tilde{Z}$ and $\tilde{U}$ are on the lines connecting 0 and Z_j and 0 and U_j correspondingly. By independence,

$$\begin{aligned} &|E[g(F_\alpha(Z^j)] - E[g(F_\alpha(Z^{j-1}))]| \\ &\leq (1/6) \sum_{m_1,m_2,m_3} \sup_{X\in\mathbb{R}^{p\times n}} \left| \frac{\partial^3 g(F_\alpha(X))}{\partial X_{jm_1}\partial X_{jm_2}\partial X_{jm_3}} \right| (E[|Z_{jm_1}Z_{jm_2}Z_{jm_3}|] + E[|U_{jm_1}U_{jm_2}U_{jm_3}|]) \end{aligned}$$

By Holder inequality,

$$E[|Z_{jm_1}Z_{jm_2}Z_{jm_3}|] \leq \max_m E[|Z_{jm}|^3] \tag{A.23}$$

and

$$E[|U_{jm_1}U_{jm_2}U_{jm_3}|] \leq \max_m E[|U_{jm}|^3] \tag{A.24}$$

Denote

$$A_1 = \sup_{X\in\mathbb{R}^{p\times n}} \left| \frac{\partial F_\alpha(X)}{\partial X_{jm_1}} \frac{\partial F_\alpha(X)}{\partial X_{jm_2}} \frac{\partial F_\alpha(X)}{\partial X_{jm_3}} \right| \tag{A.25}$$

$$\begin{aligned} A_2 = \sup_{X\in\mathbb{R}^{p\times n}} &\left| \frac{\partial F_\alpha(X)}{\partial X_{jm_1}} \frac{\partial^2 F_\alpha(X)}{\partial X_{jm_2}\partial X_{jm_3}} \right| + \\ &\sup_{X\in\mathbb{R}^{p\times n}} \left| \frac{\partial F_\alpha(X)}{\partial X_{jm_2}} \frac{\partial^2 F_\alpha(X)}{\partial X_{jm_1}\partial X_{jm_3}} \right| + \sup_{X\in\mathbb{R}^{p\times n}} \left| \frac{\partial F_\alpha(X)}{\partial X_{jm_3}} \frac{\partial^2 F_\alpha(X)}{\partial X_{jm_1}\partial X_{jm_2}} \right| \end{aligned}$$

and

$$A_3 = \sup_{X\in\mathbb{R}^{p\times n}} \left| \frac{\partial^3 F_\alpha(X)}{\partial X_{jm_1}\partial X_{jm_2}\partial X_{jm_3}} \right| \tag{A.26}$$

Then

$$\sup_{X\in\mathbb{R}^{p\times n}} \left| \frac{\partial^3 g(F_\alpha(X))}{\partial X_{jm_1}\partial X_{jm_2}\partial X_{jm_3}} \right| \leq \|g'''\|_\infty A_1 + \|g''\|_\infty A_2 + \|g'\|_\infty A_3 \tag{A.27}$$

So, it only remains to bound partial derivatives of F_α.

To simplify notation, denote $B_{km} = \exp(\alpha f_{km}(X))$ for $k = 1, ..., K$ and $m = 1, ..., p$. Then

$$\frac{\partial F_\alpha(X)}{\partial X_{jm_1}} = \frac{\sum_k B_{km_1} a_{kjm_1}}{\sum_{k,m} B_{km}} \tag{A.28}$$

The expression on the right hand side of the formula above is the expectation of a random variable which takes value a_{kjm_1} with probability $B_{km_1} / \sum_{km} B_{km}$ for $k = 1, ..., K$ and 0 with probability $1 - \sum_k B_{km_1} / \sum_{km} B_{km}$. If m_1, m_2, and m_3 are all different, then

$$\frac{\partial F_\alpha(X)}{\partial X_{jm_1}} \frac{\partial F_\alpha(X)}{\partial X_{jm_2}} \frac{\partial F_\alpha(X)}{\partial X_{jm_3}} \tag{A.29}$$

will be the product of expectations of 3 random variables with nonitersecting supports. It is easy to see that this product will be not greater than $a^3/27$. All other cases can be treated by the same argument. We have

$$A_1 \leq \begin{cases} a^3/27 & \text{if } m_1, m_2, \text{ and } m_3 \text{ are all different} \\ 4a^3/27 & \text{if } m_1 = m_2 \neq m_3 \\ a^3 & \text{if } m_1 = m_2 = m_3 \end{cases} \tag{A.30}$$

If m_1, m_2, and m_3 are all different, then

$$\frac{\partial^2 F_\alpha(X)}{\partial X_{jm_1} \partial X_{jm_2}} = -\alpha \frac{\sum_k B_{km_1} a_{kjm_1} \sum_k B_{km_2} a_{kjm_2}}{(\sum_{km} B_{km})^2} \tag{A.31}$$

and

$$\frac{\partial^3 F_\alpha(X)}{\partial X_{jm_1} \partial X_{jm_2} \partial X_{jm_3}} = 2\alpha^2 \frac{\sum_k B_{km_1} a_{kjm_1} \sum_k B_{km_2} a_{kjm_2} \sum_k B_{km_3} a_{kjm_3}}{(\sum_{km} B_{km})^3} \tag{A.32}$$

If $m_1 = m_2 \neq m_3$, then

$$\frac{\partial^2 F_\alpha(X)}{\partial X_{jm_1} \partial X_{jm_2}} = -\alpha \frac{(\sum_k B_{km_1} a_{kjm_1})^2}{(\sum_{km} B_{km})^2} + \alpha \frac{\sum_k B_{km_1} a_{kjm_1}^2}{\sum_{km} B_{km}} \tag{A.33}$$

and

$$\frac{\partial^3 F_\alpha(X)}{\partial X_{jm_1} \partial X_{jm_2} \partial X_{jm_3}}$$
$$= 2\alpha^2 \frac{(\sum_k B_{km_1} a_{kjm_1})^2 \sum_k B_{km_3} a_{kjm_3}}{(\sum_{km} B_{km})^3} - \alpha^2 \frac{\sum_k B_{km_1} a_{kjm_1}^2 \sum_k B_{km_3} a_{kjm_3}}{(\sum_{km} B_{km})^2}$$

If $m_1 = m_2 = m_3$, then

$$\frac{\partial^3 F_\alpha(X)}{\partial X_{jm_1} \partial X_{jm_2} \partial X_{jm_3}}$$
$$= \alpha^2 \frac{\sum_k B_{km_1} a_{kjm_1}^3}{(\sum_{km} B_{km})} - 3\alpha^2 \frac{\sum_k B_{km_1} a_{kjm_1}^2 \sum_k B_{km_1} a_{kjm_1}}{(\sum_{km} B_{km})^2} + 2\alpha^2 \frac{(\sum_k B_{km_1} a_{kjm_1})^3}{(\sum_{km} B_{km})^3}$$

So,

$$A_2 \leq \begin{cases} 3\alpha a^3/27 & \text{if } m_1, m_2, \text{ and } m_3 \text{ are all different} \\ 59\alpha a^3/108 & \text{if } m_1 = m_2 \neq m_3 \\ 3\alpha a^3 & \text{if } m_1 = m_2 = m_3 \end{cases} \tag{A.34}$$

and

$$A_3 \leq \begin{cases} 2\alpha^2 a^3/27 & \text{if } m_1, m_2, \text{ and } m_3 \text{ are all different} \\ 8\alpha^2 a^3/27 & \text{if } m_1 = m_2 \neq m_3 \\ \alpha^2 a^3 & \text{if } m_1 = m_2 = m_3 \end{cases} \tag{A.35}$$

Combining these bounds yields the result. □

A.4 Primitive Conditions for Assumption 1

In this section, I give a counter-example for the statement that for assumption 1 to hold, it suffices to assume that $\{X_i : i = 1, ..., n\}$ are sampled from a distribution that is absolutely continuous with respect to Lebegue measure, has bounded support, and whose density is bounded from above and away from zero on the support. I also prove that assumption 1 holds if, in addition to above conditions, one assumes that the support is a convex set.

Lemma 3. *There exists a probability distribution on $[-1, 1]^2$ that is uniform on its support such that if $\{X_i : i = 1, ..., n\}$ are sampled from this distribution, then assumption 1 fails.*

Proof. As an example of such a probability distribution, consider the uniform distribution on

$$S = \{(x_1, x_2) \in [-1, 1]^2 : x_1 \geq 0;\ -(1 + \alpha)x_1^\alpha/2 \leq x_2 \leq (1 + \alpha)x_1^\alpha/2\} \tag{A.36}$$

for some $\alpha > 0$. For fixed i, the probability that $X_{i,1} \leq \underline{h}$ is $\underline{p} = \underline{h}^{1+\alpha}$, and the probability that $X_{i,1} > \overline{h}$ is $\overline{p} = 1 - \overline{h}^{1+\alpha}$. Let A_n be an event that $X_{i,1} \leq \underline{h}$ for exactly one $i = 1, ..., n$ whereas $X_{i,1} > \overline{h}$ for all other $i = 1, ..., n$ with $\underline{h} < \overline{h}$. The probability of this event is

$$P(A_n) = n\underline{p}\overline{p}^{n-1} = n\underline{h}^{1+\alpha}(1 - \overline{h}^{1+\alpha})^{n-1} \tag{A.37}$$

Set $\underline{h} = (C_1/n)^{1/(1+\alpha)}$ and $\overline{h} = (C_2/n)^{1/(1+\alpha)}$ with $0 < C_1 < C_2 < 1$. Then I can find the

limit of $P(A_n)$ as $n \to \infty$:

$$\lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} C_1(1 - C_2/n)^{n-1} = C_1 e^{-C_2} > 0 \tag{A.38}$$

Note that on A_n, there is an observation X_i such that there is no other observations in the ball with center at X_i and radius $(C_2^{1/(1+\alpha)} - C_1^{1/(1+\alpha)})/n^{1/(1+\alpha)}$. The result now follows by choosing α sufficiently large such that $n^{-1/(1+\alpha)}$ converges to zero slower then $h_{\min}$. □

Now I give a sufficient primitive condition for assumption 1.

Lemma 4. *If $\{X_i : i = 1, ..., n\}$ are sampled from a distribution that is absolutely continuous with respect to Lebegue measure, has bounded and convex support $S \subset \mathbb{R}^d$, and whose density is bounded from above and away from zero on the support, then assumption 1 holds for sufficiently large n a.s.*

Proof. Consider sets of the following form: $I(a_1, ..., a_d, c) = S \cap \{x : a_1x_1 + ... + a_dx_d = c\}$ with $a_1^2 + ... + a_d^2 = 1$. These are convex sets. It follows from the fact that the density is bounded from above that $\inf_{a_1,...,a_d} \sup_c D(I(a_1, ..., a_d, c)) > 0$ where $D(\cdot)$ denotes the diameter of the set. So, there exists some constant $0 < C \leq 1$ such that for all $r < 1$ and all $x \in S$, each ball $B(x, r)$ with center at x and radius r has at least fraction C of its Lebegue measure inside of the support S: $\lambda(B(x, r) \cap S)/\lambda(B(x, r)) > C$.

Note that δ-covering numbers of the set S satisfy $N(\delta) \lesssim \delta^d$ as $\delta \to 0$, i.e. there exists some constant $C > 0$ such that $N(\delta) < C/\delta^d$. Consider the lower bound in assumption 1(ii). For each $h \in H_n$, consider the set of covering balls with centers $G_{h,1}$,...,$G_{h,N(h)}$ and radii $\delta_h = h/2$. Then for each X_i and $h \in H_n$, there exists some $j \in \{1, ..., N(h)\}$ such that $B(X_i, h) \supset B(G_{h,j}, \delta_h)$. Thus, it is enough to prove the lower bound for the number of observations droping into these covering balls. Since the density is bounded away from zero and from above, there exist some constants $C_1, C_2 > 0$ such that for each $h \in H_n$ and $j = 1, ..., N(h)$, $C_1h^d < P(X_i \in B(G_{h,j}, \delta_h)) < C_2h^d$. Denote $I_{h,j}(X_i) = I\{X_i \in B(G_{h,j}, \delta_h)\}$. Bernstein inequality (see proposition 1.3.2 in Dudley (1999)) gives

$$\begin{aligned} P\{\sum_{i=1}^n I_{h,j}(X_i)/n < C_1h^d/2\} &\leq P\{\sum_{i=1}^n I_{h,j}(X_i)/n - E[I_{h,d}(X_i)] < -C_1h^d/2)\} \\ &\leq C\exp(-C_1nh^d) \end{aligned}$$

Then by union bound,

$$P(\cup_{h\in H_n, j=1,...,N(h)}\{\sum_{i=1}^n I_{h,j}(X_i)/n < C_1h^d/2\}) \leq Ch_{\min}^{-d} \log n \exp(-C_1nh_{\min}^d) \tag{A.39}$$

as $n \to \infty$. By assumption 4(ii), there exists some $C > 0$ such that $nh_{\min}^d > n^C$. So, summing the probabilities above over n, I conclude, by the Borel-Cantelli lemma, that the lower bound in assumption 1(ii) holds for sufficiently large n a.s. A similar argument gives the upper bound. So, assumption 1 holds. □

A.5 Anticoncentration Inequality for the Maximum of Gaussian Random Variables

In this section, I describe an upper bound on the pdf of the maximum of correlated Gaussian random variables derived in Chernozhukov and Kengo (2011). Let $\{Z_i : i = 1, ..., S\}$ be a set of standard Gaussian (possibly correlated) random variables. Define $W = \max_{i=1,...,S} Z_i$ and let $f_W(\cdot)$ denote its pdf. Then

Lemma 5. $\sup_{w \in \mathbb{R}} f_W(w) \leq C\sqrt{\log S}$ *for some universal constant* C.

Proof. Theorem 1 in Chernozhukov and Kengo (2011) proves that $\sup_{w \in \mathbb{R}} f_W(w) \leq CE[W]$. In addition, it follows from the same argument as in lemma 8 that $E[W] \leq C\sqrt{\log n}$. Combining these bounds gives the result. □

A.6 Result on Gaussian Random Variables

In this section, I state a result on Gaussian random variables which will be used in the derivation of the lower bound on the rate of uniform consistency.

Lemma 6. *Let* ξ_n, $n = 1, ..., \infty$, *be a sequence of independent standard Gaussian random variables and* $w_{i,n}$, $i = 1, ..., n$, $n = 1, ..., \infty$, *be a triangular array of positive numbers. If* $w_{i,n} < C\sqrt{\log n}$ *with* $C \in (0, 1)$ *for all* $i = 1, ..., n$, $n = 1, ..., \infty$, *then*

$$\lim_{n \to \infty} E[|n^{-1} \sum_{i=1}^{n} \exp(w_{i,n}\xi_i - w_{i,n}^2/2) - 1|] = 0 \tag{A.40}$$

Proof. The proof is closely related to that in lemma 6.2 in Dumbgen and Spokoiny (2001). Denote $Z_{i,n} = \exp(w_{i,n}\xi_i - w_{i,n}^2/2)$ and $t_n = (E[(\sum_{i=1}^n Z_{i,n}/n - 1)^2])^{1/2}$. Note that $E[Z_{i,n}] = 1$ and $E[Z_{i,n}^2] = \exp(w_{i,n}^2)$. Thus,

$$t_n^2 = \sum_{i=1}^{n} (E[Z_{i,n}^2] - (E[Z_{i,n}])^2)/n^2 \leq \sum_{i=1}^{n} \exp(w_{i,n}^2)/n^2 \to 0 \tag{A.41}$$

if $\max_{i=1,...,n} \exp(w_{i,n}^2)/n \to 0$. The last condition holds by assumption. So,

$$\begin{aligned} E[|n^{-1}\sum_{i=1}^{n} \exp(w_{i,n}\xi_i - w_{i,n}^2/2) - 1|] &= \int_0^\infty P(|n^{-1}\sum_{i=1}^{n} Z_{i,n} - 1| > t)dt \\ &\leq t_n + \int_{t_n}^\infty t_n^2/t^2 dt \\ &= 2t_n \to 0 \end{aligned}$$

The result follows. □

A.7 Preliminary Technical Results

In this section, I derive some necessary preliminary results that are used in the proofs of the theorems stated in the main part of the paper. It is assumed throughout that assumptions 1-8 hold. I will use the following additional notation. Let $\{\psi_n\}_{n=1}^\infty$ be a sequence of positive real numbers such that $\psi_n \geq C_\psi (p \log n)^{1/2}/n^{\kappa/4}$ for some sufficiently large constant $C_\psi > 0$ and $\psi_n \to 0$ as $n \to \infty$. For any $\lambda \in (0,1)$, define $c_{1-\lambda}^{PIA,0} \in \mathbb{R}$ by analogy with $c_{1-\lambda}^{PIA}$ with Σ_i used instead of $\hat{\Sigma}_i$ for all $i = 1, ..., n$. Denote $S_n^D = \{s \in S_n : f_s/V_s > -c_{1-\gamma_n-\psi_n}^{PIA,0}\}$. For any $\lambda \in (0,1)$, define $c_{1-\lambda}^D \in \mathbb{R}$ by analogy with $c_{1-\lambda}^{RMS}$ with S_n^D used instead of S_n^{RMS}. Let $\{\epsilon_i : i = 1, ..., n\}$ be an iid sequence of p-dimensional standard Gaussian random vectors that are independent of the data. Denote $\hat{e}_j = \hat{\Sigma}^{1/2}\epsilon_j$ and $e_j = \Sigma^{1/2}\epsilon_j$. Note that $\hat{e}_j$ is equal in distribution to $\tilde{Y}_j$. Finally, denote

$$\varepsilon_{(i,m,h)} = \sum_{j=1}^{n} w_h(X_i, X_j)\varepsilon_{j,m} \tag{A.42}$$

$$f_{(i,m,h)} = \sum_{j=1}^{n} w_h(X_i, X_j) f_m(X_j) \tag{A.43}$$

$$e_{(i,m,h)} = \sum_{j=1}^{n} w_h(X_i, X_j) e_j \tag{A.44}$$

$$\hat{e}_{(i,m,h)} = \sum_{j=1}^{n} w_h(X_i, X_j) \hat{e}_j \tag{A.45}$$

$$T^{PIA} = \max_{s \in S_n} (\hat{e}_s/\hat{V}_s) \tag{A.46}$$

$$T^{PIA,0} = \max_{s \in S_n} (e_s/V_s) \tag{A.47}$$

Note that T^{PIA} is equal in distribution to the simulated statistic.

I start with a result on bounds for weights and variances of the kernel estimator. The same result can be found in Horowitz and Spokoiny (2001).

Lemma 7. *There exist constants $C > 0$ and $0 < C_1 < C_2 < \infty$ such that for any $i, j = 1, ..., n$, $m = 1, ..., p$, and $h \in H_n$,*

$$w_h(X_i, X_j) \leq C/(nh^d) \tag{A.48}$$

and

$$C_1/\sqrt{nh^d} \leq V_{(i,m,h)} \leq C_2/\sqrt{nh^d} \tag{A.49}$$

uniformly over the set of models $\mathcal{G}$.

Proof. By assumptions 1 and 7, for any $i = 1, ..., n$ and $h \in H_n$,

$$C_1 nh^d \leq CM_{h/2}(X_i) \leq \sum_{k=1}^{n} K(X_i - X_k) \leq M_h(X_i) \leq C_2 nh^d \tag{A.50}$$

and

$$C_1 nh^d \leq \sum_{k=1}^{n} K^2(X_i - X_k) \leq C_2 nh^d \tag{A.51}$$

for some constants $C > 0$ and $0 < C_1 < C_2 < \infty$. In addition, $K(X_i - X_j) \leq 1$ for any $j = 1, ..., n$. So,

$$w_h(X_i, X_j) = K(X_i - X_j)/\sum_{k=1}^{n} K(X_i - X_k) \leq C/(nh^d) \tag{A.52}$$

By assumption 2, since $\sum_{j=1}^{n} w_h(X_i, X_j) = 1$,

$$\begin{aligned} V_{(i,m,h)} &= \left(\sum_{j=1}^{n} w_h^2(X_i, X_j) \Sigma_{j,mm} \right)^{1/2} \\ &\leq C \left(\sum_{j=1}^{n} w_h^2(X_i, X_j) \right)^{1/2} \\ &\leq C \max_{j=1,...,n} w_h^{1/2}(X_i, X_j) \\ &\leq C/\sqrt{nh^d} \end{aligned}$$

and

$$V_{(i,m,h)} \geq C \left(\sum_{j=1}^{n} w_h^2(X_i, X_j) \right)^{1/2} \geq (C/nh^d) \left(\sum_{j=1}^{n} K^2(X_i - X_j) \right)^{1/2} \geq C/\sqrt{nh^d} \tag{A.53}$$

□

Lemma 8. *$E[\max_{s\in S_n}|e_s/V_s|] \leq C(\log n)^{1/2}$ uniformly over the set of models $\mathcal{G}$. In particular, $c_{1-\lambda}^{PIA,0} \leq C\sqrt{\log n}/\lambda$ for all $\lambda \in (0,1)$ uniformly over the set of models $\mathcal{G}$.*

Proof. For any $s \in S_n$, e_s/V_s is a standard Gaussian random variable. Denote $\psi = \exp(x^2)-1$. Let $\|\cdot\|_\psi$ denote ψ-Orlicz norm. It is easy to check that $\|e_s/V_s\|_\psi < C < \infty$. So, by lemma 2.2.2 in Van der Vaart and Wellner (1996),

$$E[\max_{s\in S_n}|e_s/V_s|] \leq C\|\max_{s\in S_n}|e_s/V_s|\|_\psi \leq C(\log n)^{1/2} \tag{A.54}$$

since $|S_n| \leq Cn^\phi$ for some $\phi > 0$, which gives the first result. To obtain the second result, note that Markov inequality gives

$$\lambda \leq P\{\max_{s\in S_n}|e_s/V_s| \geq c_{1-\lambda}^{PIA,0}\} \leq E[\max_{s\in S_n}|e_s/V_s|]/c_{1-\lambda}^{PIA,0} \leq C\sqrt{\log n}/c_{1-\lambda}^{PIA,0} \tag{A.55}$$

for any $\lambda \in (0,1)$. So, $c_{1-\lambda}^{PIA,0} \leq C\sqrt{\log n}/\lambda$. □

Lemma 9. *$\max_{s\in S_n}|\hat{V}_s/V_s - 1| = o_p(n^{-\kappa})$ and $\max_{s\in S_n}|V_s/\hat{V}_s - 1| = o_p(n^{-\kappa})$ uniformly over the set of models $\mathcal{G}$.*

Proof. By assumption 2, for any $(i,m,h) \in S_n$,

$$V_{(i,m,h)}^2 = \sum_{j=1}^n w_h^2(X_i,X_j)\Sigma_{j,mm} \geq C\sum_{j=1}^n w_h^2(X_i,X_j) \tag{A.56}$$

In addition,

$$|\hat{V}_{(i,m,h)}^2 - V_{(i,m,h)}^2| \leq \sum_{j=1}^n w_h^2(X_i,X_j)|\hat{\Sigma}_{j,mm} - \Sigma_{j,mm}| \tag{A.57}$$

So,

$$\begin{aligned}\max_{s\in S_n}|\hat{V}_s^2/V_s^2 - 1| &\leq C\max_{m=1,\ldots,p}\max_{j=1,\ldots,n}|\hat{\Sigma}_{j,mm} - \Sigma_{j,mm}| \\ &\leq C\max_{j=1,\ldots,n}\|\hat{\Sigma}_j - \Sigma_j\|_o\end{aligned}$$

Assumption 6 gives $\max_{j=1,\ldots,n}\|\hat{\Sigma}_j - \Sigma_j\|_o = o_p(n^{-\kappa})$. So, $\max_{s\in S_n}|\hat{V}_s^2/V_s^2 - 1| = o_p(n^{-\kappa})$. Combining this result with inequality $|x-1| \leq |x^2-1|$, which holds for any $x > 0$, yields the first result of the lemma. The second result follows from the first one and the inequality $|1/x - 1| < 2|x-1|$, which holds for any $|x-1| < 1/2$. □

Lemma 10. *$P\{c^{PIA,0}_{1-\lambda-\psi_n} > c^{PIA}_{1-\lambda}\} = o(1)$ and $P\{c^{PIA,0}_{1-\lambda+\psi_n} < c^{PIA}_{1-\lambda}\} = o(1)$ uniformly over all $\lambda \in (0,1)$[10] and over the set of models $\mathcal{G}$ where ψ_n is defined in the beginning of this section ($\psi_n \geq C_\psi (p \log n)^{1/2}/n^{\kappa/4}$ with sufficiently large $C_\psi > 0$ and $\psi_n \to 0$).*

Proof. Denote

$$p_1 = \max_{s \in S_n} \left| \frac{e_s}{V_s} \right| \max_{s \in S_n} \left| \frac{V_s}{\hat{V}_s} - 1 \right| \tag{A.58}$$

and

$$p_2 = \max_{(i,m,h) \in S_n} \left| \frac{\sum_{j=1}^n w_h(X_i, X_j)((\hat{\Sigma}_j^{1/2} - \Sigma_j^{1/2})\epsilon_j)_m}{\hat{V}_{(i,m,h)}} \right| \tag{A.59}$$

where $(\cdot)_m$ denotes m-th component of the vector $(\cdot)$. Then

$$|T^{PIA} - T^{PIA,0}| \leq p_1 + p_2 \tag{A.60}$$

Let A denote the event $\{\max_{j=1,...,n} \|\hat{\Sigma}_j - \Sigma_j\|_o < n^{-\kappa}\}$. By assumption 6, $P(A) \to 1$ as $n \to \infty$. Thus, it is enough to show that $c^{PIA,0}_{1-\lambda-\psi_n} \leq c^{PIA}_{1-\lambda}$ and $c^{PIA,0}_{1-\lambda+\psi_n} \geq c^{PIA}_{1-\lambda}$ on A.

As in the proof of lemma 9, $\max_{s \in S_n} |V_s/\hat{V}_s - 1| \leq Cn^{-\kappa}$ on A. By lemma 8, $E[\max_{s \in S_n} e_s/V_s] \leq C\sqrt{\log n}$. So, Markov inequality gives for any $B > 0$, on A,

$$P(p_1 > C\sqrt{\log n} n^{-\kappa} B | Y_1^n) \leq 1/B \tag{A.61}$$

for sufficiently large C where Y_1^n is a shorthand for $\{Y_i\}_{i=1}^n$. Consider p_2. For any $j = 1, ..., n$ and $m = 1, ..., p$,

$$\begin{aligned} E[((\hat{\Sigma}_j^{1/2} - \Sigma_j^{1/2})\epsilon_j)_m^2 | Y_1^n] &\leq E[\|(\hat{\Sigma}_j^{1/2} - \Sigma_j^{1/2})\epsilon_j\|^2 | Y_1^n] \\ &\leq E[\|\hat{\Sigma}_j^{1/2} - \Sigma_j^{1/2}\|_o^2 \|\epsilon_j\|^2 | Y_1^n] \\ &\leq p\|(\hat{\Sigma}_j^{1/2} - \Sigma_j^{1/2})\|_o^2 \\ &\leq p^2 \|\hat{\Sigma}_j - \Sigma_j\|_o \end{aligned}$$

where the last line follows from lemma 2. So, conditional on Y_1^n, on A,

$$\sum_{j=1}^n w_h(X_i, X_j)((\hat{\Sigma}_j^{1/2} - \Sigma_j^{1/2})\epsilon_j)_m / V_{(i,m,h)} \tag{A.62}$$

is a mean-zero Gaussian random variable with variance bounded by $p^2 n^{-\kappa}$ for any $(i,m,h) \in S_n$. In addition, on A, $\max_{s \in S_n} V_s/\hat{V}_s \leq 2$ for sufficiently large n. Thus, Markov inequality

[10] If $\psi_n \geq \lambda$ or $\lambda + \psi_n \geq 1$, set $c^{PIA,0}_{1-\lambda+\psi_n} = +\infty$ or $c^{PIA,0}_{1-\lambda-\psi_n} = -\infty$ correspondingly.

and the argument like that used in lemma 8 yield

$$P(p_2 > C\sqrt{\log n}pn^{-\kappa/2}B|Y_1^n) \leq 1/B \tag{A.63}$$

on A. Let $B = n^{\kappa/4}/(p\log n)^{1/2}$. Let C_1 be some large positive constant satisfying $C_1 < C_\psi$. I also assume that $C_\psi > 4$. Recall that $\psi_n \geq C_\psi(p\log n)^{1/2}/n^{\kappa/4}$. So, $\psi_n > \max(4/B,\, C_1(p\log n)n^{-\kappa/2}B)$.

Note that $T^{PIA,0}$ is the maximum over $|S_n|$ standard Gaussian random variables. Since $|S_n| \leq Cn^\phi$ for some $\phi > 0$, lemma 5 gives $c^{PIA,0}_{1-\lambda-\psi_n/2} - c^{PIA,0}_{1-\lambda-\psi_n} \geq C_2\psi_n/(\log n)^{1/2}$. I will assume that C_1 is sufficiently large so that $C_1C_2 > C$. Then

$$c^{PIA,0}_{1-\lambda-\psi_n/2} - c^{PIA,0}_{1-\lambda-\psi_n} \geq C\sqrt{\log n}pn^{-\kappa/2}B \tag{A.64}$$

Now the first part of the lemma follows from

$$\begin{aligned}
P\{T^{PIA} \leq c^{PIA,0}_{1-\lambda-\psi_n}|Y_1^n\} &\leq P\{T^{PIA,0} - p_1 - p_2 \leq c^{PIA,0}_{1-\lambda-\psi_n}|Y_1^n\} \\
&\leq P\{T^{PIA,0} - C\sqrt{\log n}pn^{-\kappa/2}B \leq c^{PIA,0}_{1-\lambda-\psi_n}|Y_1^n\} + 2/B \\
&\leq P\{T^{PIA,0} \leq c^{PIA,0}_{1-\lambda-\psi_n/2}|Y_1^n\} + 2/B \\
&\leq 1 - \lambda - \psi_n/2 + 2/B \\
&\leq 1 - \lambda
\end{aligned}$$

on A. The second part of the lemma follows from a similar argument. □

Lemma 11. *$P[\max_{s\in S_n}(\varepsilon_s/V_s) \leq c^{PIA,0}_{1-\lambda}] = 1-\lambda+o(1)$ and $E[-\max_{s\in S_n}(\varepsilon_s/V_s) \leq c^{PIA,0}_{1-\lambda}] = 1-\lambda+o(1)$ uniformly over all $\lambda \in (0,1)$ and over the set of models $\mathcal{G}$.*

Proof. By lemma 7, for any $(i,m,h) \in S_n$ and any $j = 1,...,n$,

$$w_h(X_i, X_j)/V_{(i,m,h)} \leq C/\sqrt{nh^d} \leq C/\sqrt{nh^d_{\min}} \tag{A.65}$$

It follows from assumption 4(ii) that there exists a sequence $\{\nu_n\}$ of positive numbers satisfying $\nu_n \to \infty$ sufficiently slowly so that $\nu_n^{10}(\log n)^7/(nh^{3d}_{\min}) \to 0$. Let $g_0 : \mathbb{R} \to \mathbb{R}$ be a thrice continuously differentiable function satisfying (i) $g_0(x) = 1$ for $x \leq 0$ and (ii) $g_0(x) = 0$ for $x \geq 1$. Let $g_n(x) = g_0(\nu_n\sqrt{\log n}(x - c^{PIA,0}_{1-\lambda}))$ for all $x \in \mathbb{R}$. Clearly, $\|g_n'\|_\infty \leq C\nu_n\sqrt{\log n}$, $\|g_n''\|_\infty \leq C(\nu_n\sqrt{\log n})^2$, and $\|g_n'''\|_\infty \leq C(\nu_n\sqrt{\log n})^3$. Apply theorem 6 with $g = g_n$, $Z_j = \varepsilon_j$, $Y_j = \Sigma_j^{1/2}\epsilon_j$, $a = C/\sqrt{nh^d_{\min}}$ and $K \leq Cn^\phi$ for some $\phi > 0$ and $\alpha = \nu_n^2(\log n)^{3/2}$. It follows that

$$\left| E\left[g_n(\max_{s\in S_n}(\varepsilon_s/V_s)) - g_n(\max_{s\in S_n}(e_s/V_s)) \right] \right| \to 0 \tag{A.66}$$

Therefore, the upper bound follows from

$$\begin{aligned} P\{\max_{s\in S_n}(\varepsilon_s/V_s) \le c_{1-\lambda}^{PIA,0}\} &\le E[g_n(\max_{s\in S_n}(\varepsilon_s/V_s))] \\ &\le E[g_n(\max_{s\in S_n}(e_s/V_s))] + o(1) \\ &\le P\{\max_{s\in S_n}(e_s/V_s) \le c_{1-\lambda}^{PIA,0} + 1/(\nu_n\sqrt{\log n})\} + o(1) \\ &\le P\{\max_{s\in S_n}(e_s/V_s) \le c_{1-\lambda}^{PIA,0}\} + o(1) \\ &\le 1-\lambda+o(1) \end{aligned}$$

where the last line follows from lemma 5.

The lower bound follows from the same argument with $g_n(x) = g_0(\nu\sqrt{\log n}(x - c_{1-\lambda}^{PIA,0}) + 1)$. The result for $E[-\max_{s\in S_n}(\varepsilon_s/V_s) \le c_{1-\lambda}^{PIA,0}]$ follows because $\{e_s\}_{s\in S_n}$ has a symmetric distribution. □

Lemma 12. $\max_{s\in S_n}|\varepsilon_s/V_s| = O_p(\sqrt{\log n})$ *and* $\max_{s\in S_n}|\varepsilon_s/\hat{V}_s| = O_p(\sqrt{\log n})$ *uniformly over the set of models* $\mathcal{G}$.

Proof. The result for $\max_{s\in S_n}|\varepsilon_s/V_s|$ follows from combining lemmas 8 and 11. The second result follows from

$$\max_{s\in S_n}|\varepsilon_s/\hat{V}_s| \le \max_{s\in S_n}|\varepsilon_s/V_s| \max_{s\in S_n}(V_s/\hat{V}_s) = O_p(\sqrt{\log n}) \tag{A.67}$$

since $\max_{s\in S_n}(V_s/\hat{V}_s) = O_p(1)$ by lemma 9. □

Lemma 13. $P\{\max_{s\in S_n\setminus S_n^D}\hat{f}_s/\hat{V}_s > 0\} \le o(1)$ *uniformly over the set of models* $\mathcal{G}$.

Proof. By lemma 11,

$$P\{\max_{s\in S_n}(\varepsilon_s/V_s) \le c_{1-\gamma_n-\psi_n}^{PIA,0}\} = 1-\gamma_n-\psi_n+o(1) \tag{A.68}$$

Since for any $s \in S_n\setminus S_n^D$, $f_s/V_s \le -c_{1-\gamma_n-\psi_n}^{PIA,0}$,

$$\begin{aligned} P\{\max_{s\in S_n\setminus S_n^D}(\hat{f}_s/\hat{V}_s) > 0\} &= P\{\max_{s\in S_n\setminus S_n^D}(\hat{f}_s/V_s) > 0\} \\ &= P\{\max_{s\in S_n\setminus S_n^D}(f_s/V_s + \varepsilon_s/V_s) > 0\} \\ &\le P\{\max_{s\in S_n\setminus S_n^D}(-c_{1-\gamma_n-\psi_n}^{PIA,0} + \varepsilon_s/V_s) > 0\} \\ &\le P\{\max_{s\in S_n}(\varepsilon_s/V_s) > c_{1-\gamma_n-\psi_n}^{PIA,0}\} \\ &\le 1-(1-\gamma_n-\psi_n)+o(1) \\ &= \gamma_n+\psi_n+o(1) \end{aligned}$$

Noting that $\gamma_n + \psi_n = o(1)$, which holds by the definition of ψ_n and assumption 5, yields the result. □

Lemma 14. *$P\{S_n^D \subset S_n^{RMS}\} \geq 1 + o(1)$ uniformly over the set of models $\mathcal{G}$.*

Proof. By lemma 10, $P\{c_{1-\gamma_n-\psi_n}^{PIA,0} > c_{1-\gamma_n}^{PIA}\} = o(1)$. In addition, for any $x \in (-1, 1)$,

$$2/(1+x) - 1 \geq 2(1-x) - 1 \geq 1 - 2x \geq 1 - 2|x| \tag{A.69}$$

So,

$$\begin{aligned}
P\{S_n^D \subset S_n^{RMS}\} &= P\{\min_{s\in S_n^D}(\hat{f}_s/\hat{V}_s) > -2c_{1-\gamma_n}^{PIA}\} \\
&\geq P\{\min_{s\in S_n^D}(\hat{f}_s/V_s)\max_{s\in S_n^D}(V_s/\hat{V}_s) > -2c_{1-\gamma_n}^{PIA}\} \\
&\geq P\{\min_{s\in S_n^D}(-c_{1-\gamma_n-\psi_n}^{PIA,0} + \varepsilon_s/V_s)\max_{s\in S_n^D}(V_s/\hat{V}_s) > -2c_{1-\gamma_n}^{PIA}\} \\
&= P\{\min_{s\in S_n^D}(\varepsilon_s/V_s) > c_{1-\gamma_n-\psi_n}^{PIA,0} - 2c_{1-\gamma_n}^{PIA}/\max_{s\in S_n^D}(V_s/\hat{V}_s)\} \\
&\geq P\{\max_{s\in S_n}(-\varepsilon_s/V_s) < -c_{1-\gamma_n-\psi_n}^{PIA,0} + 2c_{1-\gamma_n-\psi_n}^{PIA,0}/\max_{s\in S_n^D}(V_s/\hat{V}_s)\} + o(1) \\
&\geq P\{\max_{s\in S_n}(-\varepsilon_s/V_s) < c_{1-\gamma_n-\psi_n}^{PIA,0}(1 - 2|\max_{s\in S_n^D}(V_s/\hat{V}_s) - 1|\} + o(1)
\end{aligned}$$

By lemma 8, that $c_{1-\gamma_n-\psi_n}^{PIA,0} \leq C(\log n)^{1/2}/(\gamma_n + \psi_n)$. By lemma 9, $|\max_{s\in S_n^D}(V_s/\hat{V}_s) - 1| < Cn^{-\kappa}$ wpa1. So, wpa1,

$$c_{1-\gamma_n-\psi_n}^{PIA,0}(1 - 2|\max_{s\in S_n^D}(V_s/\hat{V}_s) - 1|) \geq c_{1-\gamma_n-\psi_n}^{PIA,0} - C(\log n)^{1/2}n^{-\kappa}/(\gamma_n + \psi_n) \tag{A.70}$$

Take $\chi_n = C(\log n)n^{-\kappa}/(\gamma_n + \psi_n)$. Then $\chi_n = o(1)$ by the choice of ψ_n. By lemma 5,

$$c_{1-\gamma_n-\psi_n}^{PIA,0} - C(\log n)^{1/2}n^{-\kappa}/(\gamma_n + \psi_n) \geq c_{1-\gamma_n-\psi_n-\chi_n}^{PIA,0} \tag{A.71}$$

Therefore,

$$\begin{aligned}
P\{S_n^D \subset S_n^{RMS}\} &\geq P\{\max_{s\in S_n}(-\varepsilon_s/V_s) < c_{1-\gamma_n-\psi_n-\chi_n}^{PIA,0}\} + o(1) \\
&\geq 1 - \gamma_n - \psi_n - \chi_n + o(1)
\end{aligned}$$

The result follows since $\gamma_n + \psi_n + \chi_n = o(1)$ by the definitions of ψ_n and χ_n and assumption 5. □

Lemma 15. *$P\{S_n^{RMS} = S_n\} \geq 1 + o(1)$ uniformly over the set of models $\mathcal{G}_{00}$.*

Proof. By lemma 10, $P\{c^{PIA,0}_{1-\gamma_n-\psi_n} > c^{PIA}_{1-\gamma_n}\} = o(1)$. By lemma 9, $\max_{s\in S_n}(V_s/\hat{V}_s) \leq 1 + n^{-\kappa}$ wpa1 as $n \to \infty$. If $f = 0_p$, then for any $s \in S_n$, $\hat{f}_s = \varepsilon_s$. So,

$$\begin{aligned}
P\{S^{RMS}_n = S_n\} &= P\{\min_{s\in S_n}(\varepsilon_s/\hat{V}_s) > -2c^{PIA}_{1-\gamma_n}\} \\
&\geq P\{\min_{s\in S_n}(\varepsilon_s/\hat{V}_s) > -2c^{PIA,0}_{1-\gamma_n-\psi_n}\} + o(1) \\
&\geq P\{\min_{s\in S_n}(\varepsilon_s/V_s)\max_{s\in S_n}(V_s/\hat{V}_s) > -2c^{PIA,0}_{1-\gamma_n-\psi_n}\} + o(1) \\
&\geq P\{\min_{s\in S_n}(\varepsilon_s/V_s)(1+n^{-\kappa}) > -2c^{PIA,0}_{1-\gamma_n-\psi_n}\} + o(1) \\
&\geq P\{\min_{s\in S_n}(\varepsilon_s/V_s) > -2c^{PIA,0}_{1-\gamma_n-\psi_n}(1-n^{-\kappa})\} + o(1) \\
&\geq P\{\min_{s\in S_n}(\varepsilon_s/V_s) > -c^{PIA,0}_{1-\gamma_n-\psi_n}\} + o(1) \\
&= P\{\max_{s\in S_n}(-\varepsilon_s/V_s) < c^{PIA,0}_{1-\gamma_n-\psi_n}\} + o(1)
\end{aligned}$$

Combining these results with lemma 11 yields

$$P\{S^{RMS}_n = S_n\} \geq 1 - \gamma_n - \psi_n + o(1) \tag{A.72}$$

The result follows by noting that $\gamma_n + \psi_n = o(1)$. □

Lemma 16. *$c^{RMS}_{1-\alpha} \leq c^{PIA}_{1-\alpha} = O_p(\sqrt{\log n})$ uniformly over the set of models $\mathcal{G}$.*

Proof. Since $S^{RMS}_n \subseteq S_n$, it follows that $c^{RMS}_{1-\alpha} \leq c^{PIA}_{1-\alpha}$. By lemma 10, $P\{c^{PIA,0}_{1-\alpha/2} < c^{PIA}_{1-\alpha}\} = o(1)$. In addition $c^{PIA,0}_{1-\alpha/2} \leq C\sqrt{\log n}$ by lemma 8. Combining these results yields the statement of the lemma. □

Lemma 17. *Let $\tau > 1$, $L > 0$, $x = (x_1, ..., x_d) \in \mathbb{R}^d$, $h = (h_1, ..., h_d) \in \mathbb{R}^d$, and $g \in \mathcal{F}_\varsigma(\tau, L)$ for some $\varsigma = 1, ..., [\tau]$. Then $\partial g(x_1, ..., x_d)/\partial x_m \geq 0$ for all $m = 1, ..., d$ implies that for any $y = (y_1, ..., y_d) \in \mathbb{R}^d$ satisfying $0 \leq y \leq h$,*

$$g(x+y) - g(x) \geq -\frac{\max(L^{\tau-[\tau]}, L)}{\prod_{j=1,...,\varsigma}(\tau - \varsigma + j)}\|h\|^\zeta \tag{A.73}$$

for $\zeta = \min(\varsigma + 1, \tau)$.

Proof. For any $y = (y_1, ..., y_d) \in \mathbb{R}^d$ satisfying $0 \leq y \leq h$, let $l = y/\|y\|$. Then $g^{(1,l)}(x) \geq 0$. If $g^{(1,l)}(x + tl) \geq 0$ for all $t \in (0, \|y\|)$, the result is obvious. If $g^{(1,l)}(x + t_0 l) = 0$ for some $t_0 \in (0, \|y\|)$, then $g^{(k,l)}(x + t_0 l) = 0$ for all $k = 1, ..., \varsigma$. If $\varsigma = [\tau]$, then by Holder smoothness, $g^{([\tau],l)}(x + tl) \geq -(L(t - t_0))^{\tau-[\tau]}$. Integrating it $[\tau]$ times gives

$$g(x+y) - g(x) \geq -\frac{L^{\tau-[\tau]}}{\prod_{j=1,...,[\tau]}(\tau - [\tau] + j)}\|y\|^\zeta \tag{A.74}$$

since $\zeta = \tau$ in this case. If $\varsigma < [\tau]$, then $g^{(\varsigma,l)}(x+tl) \geq -L(t-t_0)$. Integrating it ς times gives the inequality similar to (A.74) with $\varsigma+1$, ς, and L instead of ζ, $[\tau]$, and $L^{\tau-[\tau]}$ correspondingly. The result follows by noting that $\|y\| \leq \|h\|$. □

A.8 Proofs of Theorems

Proof of Theorem 1. Consider any $(f, \{\varepsilon_i\}_{i=1}^{\infty}, \{X_i\}_{i=1}^{\infty}) \in \mathcal{G}_0$. For any $s \in S_n$, $f_s \leq 0$ since the kernel K is positive by assumption 7. By lemma 10, $P\{c_{1-\alpha-\psi_n}^{PIA,0} > c_{1-\alpha}^{PIA}\} = o(1)$. By lemma 9, $\max_{s\in S_n}(V_s/\hat{V}_s) \leq 1 + n^{-\kappa}$ wpa1 as $n \to \infty$. So,

$$\begin{aligned}
P\{\hat{T} \leq c_{1-\alpha}^{PIA}\} &= P\{\max_{s\in S_n}(\hat{f}_s/\hat{V}_s) \leq c_{1-\alpha}^{PIA}\} \\
&\geq P\{\max_{s\in S_n}(\varepsilon_s/\hat{V}_s) \leq c_{1-\alpha}^{PIA}\} \\
&\geq P\{\max_{s\in S_n}(\varepsilon_s/\hat{V}_s) \leq c_{1-\alpha-\psi_n}^{PIA,0}\} + o(1) \\
&\geq P\{\max_{s\in S_n}(\varepsilon_s/V_s) \max_{s\in S_n}(V_s/\hat{V}_s) \leq c_{1-\alpha-\psi_n}^{PIA,0}\} + o(1) \\
&\geq P\{\max_{s\in S_n}(\varepsilon_s/V_s)(1+n^{-\kappa}) \leq c_{1-\alpha-\psi_n}^{PIA,0}\} + o(1)
\end{aligned}$$

Let $\chi_n = (\log n)^{3/2} n^{-\kappa}$. Since $\max_{s\in S_n} |\varepsilon_s/V_s| = O_p(\sqrt{\log n})$ by lemma 12, an application of lemma 5 shows that the last expression is bounded from below by

$$P\{\max_{s\in S_n}(\varepsilon_s/V_s) \leq c_{1-\alpha-\psi_n-\chi_n}^{PIA,0}\} + o(1) \tag{A.75}$$

Then $P\{\hat{T} \leq c_{1-\alpha}^{PIA}\} \geq 1 - \alpha + o(1)$ follows from this bound and lemma 11 since $\psi_n + \chi_n \to 0$.

Now consider the RMS critical value. By lemma 14, $P\{c_{1-\alpha}^{D} > c_{1-\alpha}^{RMS}\} \leq o(1)$. By lemma 13, $P\{\max_{s\in S_n \setminus S_n^D} \hat{f}_s/\hat{V}_s > 0\} \leq o(1)$. So,

$$\begin{aligned}
P\{\hat{T} \leq c_{1-\alpha}^{RMS}\} &= P\{\max_{s\in S_n}(\hat{f}_s/\hat{V}_s) \leq c_{1-\alpha}^{RMS}\} \\
&\geq P\{\max_{s\in S_n}(\hat{f}_s/\hat{V}_s) \leq c_{1-\alpha}^{D}\} + o(1) \\
&\geq P\{\max_{s\in S_n^D}(\hat{f}_s/\hat{V}_s) \leq c_{1-\alpha}^{D}\} + o(1)
\end{aligned}$$

Since S_n^D is nonstochastic, from this point, the argument similar to that used in the proof for the plug-in test function with S_n^D instead of S_n yields the result for the RMS critical value. Note that all asymptotic results in this part of the proof hold uniformly over $\mathcal{G}_0$.

Next consider any $(f, \{\varepsilon_i\}_{i=1}^{\infty}, \{X_i\}_{i=1}^{\infty}) \in \mathcal{G}_{00}$ so that $f = 0_p$. By lemma 10, $P\{c_{1-\alpha+\psi_n}^{PIA,0} <$

$c_{1-\alpha}^{PIA}\} = o(1)$. By lemma 9, $\min_{s\in S_n}(V_s/\hat{V}_s) \geq 1 - n^{-\kappa}$ wpa1 as $n \to \infty$. So,

$$\begin{aligned} P\{\hat{T} \leq c_{1-\alpha}^{PIA}\} &= P\{\max_{s\in S_n}(\hat{f}_s/\hat{V}_s) \leq c_{1-\alpha}^{PIA}\} \\ &= P\{\max_{s\in S_n}(\varepsilon_s/\hat{V}_s) \leq c_{1-\alpha}^{PIA}\} \\ &\leq P\{\max_{s\in S_n}(\varepsilon_s/\hat{V}_s) \leq c_{1-\alpha+\psi_n}^{PIA,0}\} + o(1) \\ &\leq P\{\max_{s\in S_n}(\varepsilon_s/V_s) \min_{s\in S_n}(V_s/\hat{V}_s) \leq c_{1-\alpha+\psi_n}^{PIA,0}\} + o(1) \\ &\leq P\{\max_{s\in S_n}(\varepsilon_s/V_s)(1 - n^{-\kappa}) \leq c_{1-\alpha+\psi_n}^{PIA,0}\} + o(1) \end{aligned}$$

An argument like that used above shows that the last expression equals $1 - \alpha + o(1)$.

For the RMS critical value, note that by lemma 15, $P\{S_n^{RMS} = S_n\} \geq 1 + o(1)$ whenever $f = 0_p$. So,

$$P\{\hat{T} \leq c_{1-\alpha}^{RMS}\} = P\{\hat{T} \leq c_{1-\alpha}^{PIA}\} + o(1) = 1 - \alpha + o(1) \tag{A.76}$$

Note that all asymptotic results in this part of the proof hold uniformly over $\mathcal{G}_{00}$. □

Proof of Theorem 2. For any $w = (f, \{\varepsilon_i\}_{i=1}^{\infty}, \{X_i\}_{i=1}^{\infty}\} \in \mathcal{G}_\rho$, there exist $i \in \mathbb{N}$ and $m = 1, ..., p$ such that $f_m(X_i) \geq 3\rho/4$. By assumption 3, there exists a ball $B_\delta(X_i)$ with center at X_i and radius δ such that $f_m(X_j) \geq \rho/2$ for all $X_j \in B_\delta(X_i)$. Note that δ can be chosen independently of w. So, for some $N \in \mathbb{N}$ and any $n \geq N$, there exists a triple $s_n = (i_n, m, h_n) \in S_n$ with h_n bounded away from zero such that $f_m(X_j) \geq \rho/2$ for all $X_j \in B_{h_n}(X_{i_n})$. Hence, $f_{s_n} \geq \rho/2$. Lemma 7 gives $V_{s_n} \leq n^{-\phi}$ for some $\phi > 0$, so $f_{s_n}/V_{s_n} > Cn^{\phi}$. By lemma 9, $|\hat{V}_{s_n}/V_{s_n} - 1| = o_p(1)$. So, for any $\tilde{C} < C$, $P\{f_{s_n}/\hat{V}_{s_n} > \tilde{C}n^{\phi}\} \to 1$. Thus,

$$\begin{aligned} P\{\hat{T} \leq c_{1-\alpha}^{P}\} &\leq P\{f_{s_n}/\hat{V}_{s_n} \leq c_{1-\alpha}^{P} + \max_{s\in S_n}|\varepsilon_s/\hat{V}_s|\} \\ &\leq P\{c_{1-\alpha}^{P} + \max_{s\in S_n}|\varepsilon_s/\hat{V}_s| > \tilde{C}n^{\phi}\} + o(1) \end{aligned}$$

The result follows by noting that from lemmas 12 and 16, $c_{1-\alpha}^{P} + \max_{s\in S_n}|\varepsilon_s/\hat{V}_s| = O_p(\sqrt{\log n})$. □

Proof of Theorem 3. As in the proof of theorem 2, since $\rho(w(0), H_0) > 0$, there exists $i \in \mathbb{N}$ such that $f_m^0(X_i) \geq \rho$ for some $m = 1, ..., p$ and $\rho > 0$. In addition, by assumption 3, there exists a ball $B_\delta(X_i)$ such that $f_m^0(X_j) \geq \rho/2$ for all $X_j \in B_\delta(X_i)$. So, for some $N \in \mathbb{N}$ and any $n \geq N$, there exists a triple $s_n = (i_n, m, h) \in S_n$ such that $f_m^0(X_j) \geq \rho/2$ for all $X_j \in B_h(X_{i_n})$. Hence, $f_{s_n}^n \geq a_n\rho/2$. By lemma 7, $V_{s_n} \leq C/\sqrt{n}$. Then lemma 9 gives $P\{f_{s_n}^n/\hat{V}_{s_n} > \tilde{C}a_n/\sqrt{n}\} \to 1$ for some $\tilde{C} > 0$. The same argument as in the proof of theorem

2 yields

$$P\{\hat{T} \leq c^P_{1-\alpha}\} \leq P\{c^P_{1-\alpha} + \max_{s \in S_n} |\varepsilon_s / \hat{V}_s| > \tilde{C} a_n \sqrt{n}\} + o(1) \tag{A.77}$$

Combining $c^P_{1-\alpha} + \max_{s \in S_n} |\varepsilon_s / \hat{V}_s| = O_p(\sqrt{\log n})$ and $a_n \sqrt{n / \log n} \to \infty$ gives the result. □

Proof of Theorem 4. First, consider $\tau \leq 1$ case. In this case, $\zeta = \tau$. Since $d \geq 1$, I have $\zeta \leq d$. For any $w = (f, \{\varepsilon_i\}_{i=1}^\infty, \{X_i\}_{i=1}^\infty) \in \mathcal{G}_\vartheta$, there exist $i \in \mathbb{N}_\vartheta$ and $m = 1, ..., p$ such that $f_m(X_i) \geq (C/2) h_{\min}^\zeta$. By assumptions 1 and 8, there exists $j = 1, ..., n$ such that $\|X_i - X_j\| \leq 3 h_{\min}$ and $s_n(w) = (j, m, h_{\min}) \in S_n$. By assumption 3, $f_m(X_l) \geq \tilde{C} h_{\min}^\zeta$ for all $l = 1, ..., n$ such that $X_l \in B_{h_{\min}}(X_j)$ for some constant $\tilde{C}$. So, $f_{s_n(w)} \geq \tilde{C} h_{\min}^\zeta$. By assumption 4(ii), $n h_{\min}^{3d} / \log n \to \infty$ as $n \to \infty$. By lemma 7, $V_{s_n(w)} \leq C / \sqrt{n h_{\min}^d}$. So,

$$f_{s_n(w)} / (V_{s_n(w)} \sqrt{\log n}) \geq (\tilde{C}/C) \sqrt{n h_{\min}^{2\zeta + d} / \log n} \geq (\tilde{C}/C) \sqrt{n h_{\min}^{3d} / \log n} \to \infty \tag{A.78}$$

uniformly over $w \in \mathcal{G}_\vartheta$. The result follows from the same argument as in the proof of theorem 2.

Consider $\tau > 1$ case. Suppose $\zeta \leq d$. For any $w = (f, \{\varepsilon_i\}_{i=1}^\infty, \{X_i\}_{i=1}^\infty) \in \mathcal{G}_\vartheta$, there exist $i \in \mathbb{N}_\vartheta$ and $m = 1, ..., p$ such that $f_m(X_i) \geq (C/2) h_{\min}^\zeta$. For $m = 1, ..., d$, set $e_m = 4 h_{\min}$ if $\partial f_m(X_i) / \partial x_m \geq 0$ and $-4 h_{\min}$ otherwise. Consider the cube $\mathcal{C}$ whose edges are parallel to axes and that contains vertices $(X_{i,1}, ..., X_{i,d})$ and $(X_{i,1} + 2e_1, ..., X_{i,d} + 2e_d)$. By lemma 17, for all $x \in \mathcal{C}$, $f_m(x) \geq \tilde{C} h_{\min}^\zeta$ for some constant $\tilde{C}$. By the definition of $\mathbb{N}_\vartheta$ and assumption 1, there exists $l = 1, ..., n$ such that $X_l \in B_{h_{\min}}(X_{i,1} + e_1, ..., X_{i,d} + e_d)$. By assumption 8, there exists $j = 1, ..., n$ such that $X_j \in B_{3h_{\min}}(X_{i,1} + e_1, ..., X_{i,d} + e_d)$ and $s_n(w) = (j, m, h_{\min}) \in S_n$. So, $f_m(X_l) \geq \tilde{C} h_{\min}^\zeta$ for all $l = 1, ..., n$ such that $X_l \in B_{h_{\min}}(X_j)$. The rest of the proof follows from the same argument as in the case $\tau \leq 1$.

Suppose $\zeta > d$. The only difference between this case and the previous one is that now optimal testing bandwidth value is greater than $h_{\min}$. Let h_o be the largest bandwidth value in the set S_n that is smaller than $(\log n / n)^{1/(2\zeta + d)}$. For any $w \in \mathcal{G}_\vartheta$, the same construction as above gives $s_n(w) = (j, m, h_o) \in S_n$ such that $f_m(X_l) \geq \rho_\vartheta(w, H_0) - \tilde{C} h_o^\zeta$ for all $l = 1, ..., n$ such that $X_l \in B_{h_o}(X_j)$. Since $\rho_\vartheta(w, H_0) \geq b_n (\log n / n)^{\zeta/(2\zeta + d)}$ for some sequence of real numbers $\{b_n\}_{n=1}^\infty$ such that $b_n \to \infty$ as $n \to \infty$, $f_{s_n(w)} \geq (b_n - \tilde{C})(\log n / n)^{\zeta/(2\zeta + d)}$. By lemma 7, $V_{s_n(w)} \leq C / \sqrt{n h_o^d}$. Then

$$f_{s_n(w)} / (V_{s_n(w)} \sqrt{\log n}) \geq (b_n - \tilde{C}) / (2C) \to \infty \tag{A.79}$$

The result follows as above. □

Proof of Theorem 5. First, define functions $b_1, ..., b_K$ on $(0, 1]$ for $K = [\tau]$ by the following

induction. Set $b_1(x) = +1$ for $x \in (0, 1/2]$ and -1 for $x \in (1/2, 1]$. Given $b_1, ..., b_{k-1}$, for $i = 1, 3, ..., 2^k - 1$ and $x \in ((i-1)2^{-k}, i2^{-k}]$, set $b_k(x) = +1$ if $b_{k-1}(y) = +1$ for $y \in ((i-1)2^{-k}, (i+1)2^{-k}]$ and -1 otherwise. For $i = 2, 4, ..., 2^k$ and $x \in ((i-1)2^{-k}, i2^{-k}]$, set $b_k(x) = -1$ if $b_{k-1}(y) = +1$ for $y \in ((i-2)2^{-k}, i2^{-k}]$ and $+1$ otherwise.

Now let us define $v : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$. Set $v(x, h) = 0$ if $x < 0$ or $x > 2$ for all $h \in \mathbb{R}_+$. For $x \in [0, 2]$, v will be defined through its derivatives. Set $\partial^k v(0, h)/\partial x^k = 0$ for all $k = 0, ..., K$. For $i = 1, ..., 2^K$, once function $\partial^K v(x, h)/\partial x^K$ is defined for $x \in [0, (i-1)2^{-K}]$, set

$$\partial^K v(x, h)/\partial x^K = \partial^K v((i-1)2^{-K}, h)/\partial x^K + b_K(x) h^K L(x - (i-1)2^{-K})^{\tau - K} \tag{A.80}$$

for $x \in ((i-1)2^{-K}, i2^{-K}]$. These conditions define function $v(x, h)$ for $x \in [0, 1]$ and $h \in \mathbb{R}_+$. For $x \in (1, 2]$ and $h \in \mathbb{R}_+$, set $v(x, h) = v(2-x, h)$ so that v is symmetric in x around $x = 1$. It is easy to see that for fixed $h \in \mathbb{R}_+$, $v(\cdot/h, h) \in \mathcal{F}_{[\tau]}(\tau, L)$ and $\sup_{x \in \mathbb{R}} v(x/h, h) \in (C_1 h^\tau, C_2 h^\tau)$ for some positive constants C_1 and C_2 independent of h.

Let $q : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}_+$ be given by $q(x, h) = v(\|x\|/h + 1, h)$ for all $(x, h) \in \mathbb{R}^d \times \mathbb{R}_+$. Note that for fixed $h \in \mathbb{R}_+$, $q(\cdot, h) \in \mathcal{F}_{[\tau]}(\tau, L)$, $q(x, h) = 0$ if $\|x\| > h$, and $q(0_d, h) = \sup_{x \in \mathbb{R}^d} q(x, h) \in (C_1 h^\tau, C_2 h^\tau)$.

Since $r_n(n/\log n)^{\tau/(2\tau+d)} \to 0$, there exists a sequence of positive numbers $\{\psi_n\}_{n=1}^\infty$ such that $r_n = \psi_n^\tau (\log n/n)^{\tau/(2\tau+d)}$ and $\psi_n \to 0$. Set $h_n = \psi_n (\log n/n)^{1/(2\tau+d)}$. By the assumption on packing numbers $N(h, S_\vartheta)$, there exists a set $\{j(l) \in \mathbb{N}_\vartheta : l = 1, ..., N_n\}$ such that $\|X_{j(l_1)} - X_{j(l_2)}\| > 2h_n$ for $l_1, l_2 = 1, ..., N_n$ if $l_1 \neq l_2$ and $N_n > C h_n^{-d}$ for some constant C. For $l = 1, ..., N_n$, define function $f^l : \mathbb{R}^d \to \mathbb{R}^p$ given by $f_1^l(x) = q(x - X_{j(l)}, h_n)$ and $f_m^l(x) = 0$ for all $m = 2, ..., p$ for all $x \in \mathbb{R}^d$. Note that functions $\{f^l\}_{l=1}^{N_n}$ have disjoint supports. Moreover, for every $l = 1, ..., N_n$ and $m = 1, ..., p$, $f_m^l \in \mathcal{F}_{[\tau]}(\tau, L)$. Let $\{\varepsilon_i\}_{i=1}^\infty$ be a sequence of independent standard Gaussian random vectors $N(0, I_p)$. For $l = 1, ..., N_n$, define an alternative $w_l = (f^l, \{\varepsilon_i\}_{i=1}^\infty, \{X_i\}_{i=1}^\infty\}$. Note that $\rho_\vartheta(w_l, H_0) \geq C r_n$ for all $l = 1, ..., N_n$ for some constant C. In addition, let $w_0 = (0, \{\varepsilon_i\}_{i=1}^\infty, \{X_i\}_{i=1}^\infty)$.

As in the proof of lemma 6.2 in Dumbgen and Spokoiny (2001), for any sequence $\phi_n =$

$\phi_n(Y_1, ..., Y_n)$ of tests with $\sup_{w \in \mathcal{G}_0 \cap \mathcal{G}_X} E_w[\phi_n] \leq \alpha$,

$$\begin{aligned}
\inf_{w \in \mathcal{G}_X, \rho_\vartheta(w, H_0) \geq C r_n} E_w[\phi_n] - \alpha &\leq \min_{l=1,...,N_n} E_{w_l}[\phi_n] - E_{w_0}[\phi_n] \\
&\leq \sum_{i=1}^{N_n} E_{w_l}[\phi_n]/N_n - E_{w_0}[\phi_n] \\
&\leq E_{w_0}[(\sum_{i=1}^{N_n} (dP_{w_l}/dP_{w_0})/N_n - 1)\phi_n] \\
&\leq E_{w_0}[|\sum_{i=1}^{N_n} dP_{w_l}/dP_{w_0}/N_n - 1|]
\end{aligned}$$

where dP_{w_l}/dP_{w_0} denotes a Radon-Nykodim derivative. For $l = 1, ..., N_n$, denote $\omega_l = (\sum_{i=1}^n (f_1^l(X_i))^2)^{1/2}$ and $\xi_l = \sum_{i=1}^n f_1^l(X_i)\varepsilon_{i,1}/\omega_l$. Then

$$dP_{w_l}/dP_{w_0} = \exp(\omega_l \xi_l - \omega_l^2/2) \tag{A.81}$$

Note that $\omega_l \leq C n^{1/2} h_n^{\tau + d/2}$. In addition, under the model w_0, ξ_l are independent standard Gaussian random variables. So, an application of lemma 6 gives

$$E_{w_0}[|\sum_{i=1}^{N_n} dP_{w_l}/dP_{w_0}/N_n - 1|] \to 0 \tag{A.82}$$

if $C n^{1/2} h_n^{\tau + d/2} < \tilde{C} (\log N_n)^{1/2}$ for some constant $\tilde{C} \in (0, 1)$ for all large enough n. The result follows by noting that $n^{1/2} h_n^{\tau + d/2} = o(\sqrt{\log n})$ and $\log N_n \geq C \log n$ for some constant C. □

Proof of Corollary 1. Replace p by Q_n everywhere in the proofs given above. Then all preliminary results except lemma 11 hold for the test with $Q_n \to \infty$. Lemma 11 holds with condition (iv) in the corollary replacing assumption 4(ii). So, the first result follows from the same argument as in theorem 1. For any $w \in \mathcal{G}_\rho$, there exists some $m \in \mathbb{N}$ such that $\sup_{i \in \mathbb{N}} [f_m(X_i)]_+ > 0$. Once m is included in the test statistic, the second result follows as in the proof of theorem 2. □

Proof of Corollary 2. To prove the first result, note that $f_m^{x,z}(X_i, Z_i) < Ca$ for all $i \in N$ and $m = 1, ..., p$ by assumption (v). So, $f_s \leq Ca$ for any $s \in S_n$. Therefore, combining lemmas 7 and 9 gives

$$\max_{s \in S_n} (f_s / \hat{V}_s) \leq Ca \sqrt{n_a h_{\max}^d}$$

wpa1. Since $a \sqrt{n_a h_{\max}^d \log n} \to 0$ by assumption (iv), the bias is asymptotically negligible

in comparison with the concentration rate of the test statistic. Therefore, the argument like that used in the proof of theorem 1 leads to

$$P\{\hat{T} \leq c^P_{1-\alpha}\} \geq P\{\max_{s \in S_n}(\varepsilon^{x,z}_s / V_s) \leq c^P_{1-\alpha}\} + o(1) = 1 - \alpha + o(1) \tag{A.83}$$

as $n \to \infty$ for $P = PIA$ or RMS.

The second result follows from the same argument since under assumptions (v) and (vi) $-Ca \leq f_s \leq Ca$.

Finally, consider the third part of the corollary. If $\rho_z(w, H_0) > \rho$, then for sufficiently large n, there exists a triple $s_n = (i_n, m, h_n) \in S_n$ with h_n bounded away from zero such that $f_m(X_j) \geq \rho/2$ for all $X_j \in B_{h_n}(X_{i_n})$ and $\|Z_{i_n} - z_0\| \leq a_n$. The rest of the proof follows from the argument similar to that used in the proof of theorem 2. □

References

Andrews, D. W. K., Guggenberger, P., 2009. Validity of subsampling and plug-in asymptotic inference for parameters defined by moment inequalities. Econometric Theory 25, 669–709.

Andrews, D. W. K., Han, S., 2009. Invalidity of the bootstrap and m out of n bootstrap for interval endpoints. Econometrics Journal 12, S172–S199.

Andrews, D. W. K., Shi, X., 2010. Inference based on conditional moment inequalities. Cowles Foundation Discussion Paper, No 1761.

Andrews, D. W. K., Soares, G., 2010. Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica 78, 119–157.

Armstrong, T., 2011a. Asymptotically exact inference in conditional moment inequalities models. unpublished manuscript.

Armstrong, T., 2011b. Weighted ks statistics for inference on conditional moment inequalities. unpublished manuscript.

Bugni, F. A., 2010. Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. Econometrica 78, 735–753.

Canay, I. A., 2010. El inference for partially identified models: Large deviations optimality and bootstrap validity. Journal of Econometrics 156, 408–425.

Chatterjee, S., 2005. A simple invariance theorem. arXiv:math/0508213v1.

Chernozhukov, V., Hong, H., Tamer, E., 2007. Estimation and confidence regions for parameter sets in econometric models. Econometrica 75, 1243–1284.

Chernozhukov, V., Kengo, K., 2011. Anti-concentration and honest adaptive confidence bands. working paper, 1–20.

Chernozhukov, V., Lee, S., Rosen, A. M., 2009. Intersection bounds: Estimation and inference. CEMMAP working paper CWP 19/09.

Dudley, R., 1999. Uniform Central Limit Theorems. Cambridge Studies in Advanced Mathematics.

Dumbgen, L., Spokoiny, V. G., 2001. Multiscale testing of qualitative hypotheses. The Annals of Statistics 29, 124–152.

Fan, J., Yao, Q., 1998. Efficient estimation of conditional variance functions in stochastic regression. Biometrica 85, 645–660.

Haile, P., Tamer, E., 2003. Inference with an incomplete model of english auctions. Journal of Political Economy 111, 1–51.

Hardle, W., Tsybakov, A., 1997. Local polinomial estimators of the volatility function in nonparametric autoregression. Journal of Econometrics 81, 233–242.

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

Khan, S., Tamer, E., 2009. Inference on endogenously censored regression models using conditional moment inequalities. Journal of Econometrics 152, 104–119.

Kim, K., 2008. Set estimation and inference with models characterized by conditional moment inequalities. unpublished manuscript, University of Minnesota.

Kress, R., 1999. Linear Integral Equations. Springer.

Lee, S., Song, K., Whang, Y. J., 2011. Testing function inequalities. CEMMAP working paper CWP 12/11.

Lepski, O. V., Spokoiny, V. G., 1999. Minimax nonparametric hypothesis testing: the case of an inhomogeneous alter. Bernoulli 5, 333–358.

Manski, C. F., Tamer, E., 2002. Inference on regressions with interval data on a regressor or outcome. Econometrica 70, 519–546.

Milgrom, P., Weber, R., 1982. A theory of auctions and competitive bidding. Econometrica 50, 1089–1122.

Muller, H. G., Stadtmuller, U., 1987. Estimation of heteroscedasticity in regression analysis. The Annals of Statistics 15, 610–625.

Pakes, A., 2010. Alternative models for moment inequalities. Econometrica 78, 1783–1822.

Rice, J., 1984. Bandwidth choice for nonparametric kernel regression. The Annals of Statistics 12, 1215–1230.

Romano, J., P., Shaikh, A. M., 2008. Inference for identifiable parameters in partially identified econometric models. Journal of Statistical Planning and Inference 138, 2786–2807.

Romano, J., P., Shaikh, A. M., 2010. Inference for the identified sets in partially identified econometric models. Econometrica 78, 169–211.

Rosen, A. M., 2008. Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities. Journal of Econometrics 146, 107–117.

Tsybakov, A., 2009. Introduction to Nonparametric Estimation. Springer.

Van der Vaart, A. W., Wellner, J. A., 1996. Weak Convergence and Empirical Processes with Applications to Statistics. Springer.