ADAPTIVE TEST OF CONDITIONAL MOMENT INEQUALITIES

DENIS CHETVERIKOV

Abstract. Many economic models yield conditional moment inequalities that can be used for inference on parameters of these models. 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 a 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-behaved alternatives against which the test developed in this paper is consistent and those tests are not.

1. Introduction

Conditional moment inequalities (CMI) are important both in economics and in econometrics. In economics, they arise naturally in many models that include behavioral choice; see [31] for a survey. In econometrics, they appear in estimation problems with interval data and problems with censoring; see, for example, [28]. In addition, CMI offer a convenient way to study treatment effects in randomized experiments as described in [26]. In the next section, I provide three detailed examples of models with CMI.

To describe CMI model, 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. Then CMI are given by the following inequalities:

$$E[m_j(X, W, \theta)|X] \leqslant 0 \text{ for all } j = 1, ..., p \, a.s. \tag{1}$$

where $m_j(\cdot, \cdot, \cdot)$ is the jth component of the vector-valued function $m(\cdot, \cdot, \cdot)$. Note that (1) also covers conditional moment equalities (CME) because CME can be represented as pairs of CMI. In this paper, 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).

Date: First version: November 2010. This version: November 3, 2013. Email: chetverikov@econ.ucla.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, Simon Lee, Rosa Matzkin, Anna Mikusheva, and Adam Rosen for useful comments and discussions. The first version of the paper was presented at the Econometric lunch at MIT on November 18, 2010.

The main goal of this paper is to derive a test that has good power simultaneously for many different functions $m(\cdot,\cdot,\cdot)$. In fact, the derived test will be shown to be optimal in a sense defined below. To motivate the test, let $\rho(\cdot,\cdot)$ be a metric on the space of functions $f:\mathbb{R}^d\to\mathbb{R}^p$ and let $\mathcal{M}$ denote the set of functions $f(\cdot)$ satisfying $f_j(X)\leqslant 0$ for all $j=1,...,p$ a.s. Then the power of any test of H_0 against H_a depends on how far $E[m(X,W,\theta_0)|X=\cdot]$ is from $\mathcal{M}$, i.e. how large

$$\rho=\rho\left(E[m(X,W,\theta_0)|X=\cdot],\mathcal{M}\right)=\inf_{f\in\mathcal{M}}\rho\left(E[m(X,W,\theta_0)|X=\cdot],f(\cdot)\right)$$

is. For fixed ρ, the power of the test typically depends on the form of $E[m(X,W,\theta_0)|X=\cdot]$. Note, however, that $E[m(X,W,\theta_0)|X=\cdot]$ is a (vector) of *nonparametric* functions. Therefore, to obtain a test with good power for many different functions $m(\cdot,\cdot,\cdot)$, I consider alternatives with $E[m(X,W,\theta_0)|X=\cdot]$ belonging to large nonparametric classes of functions. Further, several concepts can be used to assess whether the test has good power. In this paper, I employ the concept of *rate optimality* or, more precisely, the concept of *minimax rate optimality*. To define this concept, for $L>0$ and $\tau\in(0,1]$, let $\mathcal{F}(\tau,L)$ be a class of functions $f:\mathbb{R}^d\to\mathbb{R}^p$ satisfying $|f_j(x)-f_j(y)|\leqslant L|x-y|^\tau$ for all $j=1,...,p$, and let ρ_n be a sequence of constants converging to zero. Then the (minimax) optimal rate is the fastest rate of convergence of ρ_n to zero such that there exists a test with the correct asymptotic size and that is uniformly consistent against the class of models satisfying $E[m(X,W,\theta_0)|X=\cdot]\in\mathcal{F}(\tau,L)$ and $\rho(E[m(X,W,\theta_0)|X=\cdot],\mathcal{M})\geqslant\rho_n$. The corresponding test is called (minimax) rate optimal. Thus, the test is (minimax) rate optimal if there does not exist another test that would control asymptotic size and would have a faster rate of uniform consistency. The test developed in this paper will be (minimax) rate optimal for all $\tau\in(0,1]$ when $\rho(\cdot,\cdot)$ is the uniform metric on the space of functions $f:\mathbb{R}^d\to\mathbb{R}^p$.[1] In addition, implementing the test does not require the knowledge of the smoothness parameter τ. The test automatically adapts to the function class where the function $E[m(X,W,\theta_0)|X=\cdot]$ belongs to and detects alternatives in this class with the optimal rate. In this sense, the test is called adaptive. To the best of my knowledge, this paper is the first to provide a test with such optimality properties; see the literature review below. The adaptivity of the test is important because smoothness properties of moment functions are rarely known in practice.

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

$$\Theta_I=\{\theta\in\Theta:\ E[m_j(X,W,\theta)|X]\leqslant 0\text{ for all }j=1,...,p\ a.s.\}$$

denote the identified set. CMI are said to identify θ if and only if Θ_I is a singleton. Otherwise, CMI do not identify the parameter θ. For example, non-identification may happen when the

[1]The test in this paper will be rate optimal in classes $\mathcal{F}(\tau,L)$ with $\tau>1$ as well. Note, however, that when $\tau>1$, classes $\mathcal{F}(\tau,L)$ must be defined differently because $|f_j(x)-f_j(y)|\leqslant L|x-y|^\tau$ for all x and y in this case implies that f is identically constant. Appropriate definitions are given in Section 4 below.

CMI arise from a game-theoretic model with multiple equilibria. Moreover, the parameter may be weakly identified, which means that Θ_I is a singleton but information on Θ_I contained in the data is limited even in large samples. My approach leads to a robust test with the correct asymptotic size no matter whether the parameter θ is identified, weakly identified, or not identified. Here "robust" means that the test controls asymptotic size uniformly over a large class of models.

The test statistic in this paper is based on kernel estimates of conditional moment functions $E[m_j(X, W, \theta_0)|X]$ with many different bandwidth values. 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 bounded away from zero. Since the variance of the kernel estimators varies greatly with the bandwidth value, each estimate is studentized. In other words, each estimate is divided by its estimated standard deviation. The test statistic, $\widehat{T}$, is formed as the maximum of these studentized estimates, and large values of $\widehat{T}$ 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 in large samples depends on the distribution of the noise $\{m(X_i, W_i, \theta_0) - E[m(X_i, W_i, \theta_0)|X_i]\}_{i=1}^n$ only via second moments of the noise. For reasons similar to those discussed in [11] and [5], 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. Therefore, 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 [3], I use a model-specific, data-driven, critical value for the pretest, which is taken to be a large quantile of the appropriate distribution. I also provide a plug-in critical value for the test. My proof of the bootstrap validity is nonstandard because it uses only finite sample arguments. My proof is also different from those used in [3] and [14].

One of the advantages of the proof technique used in this paper is that it gives an explicit error bound on the bootstrap approximation error of the distribution of the test statistic. In particular, it allows me to show that for the test developed in this paper, the probability of rejecting the null under the null can exceed nominal level only by a *polynomially* (in n) small term. In contrast, all other papers on CMI only show that this probability is *asymptotically* not larger than the nominal level. I believe that this contribution is important in light of the fact that asymptotic approximations of suprema of processes typically provide only *logarithmically* (in n) small approximation error; see, for example, [20].

Testing moment inequalities has been a popular research topic in econometrics recently. As a result, the literature concerned with unconditional and conditional moment inequalities is

expanding quickly. Published papers on unconditional moment inequalities include [11], [34], [35], [1], [2], [5], [9], [10], [31], and [33]. 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 [23], [24], [14], [3], [26], [6], [7], and [8]. The case of CMI that point identify θ is treated in [23]. All other papers from the list above deal with partial identification. The method of [3] is based on converting CMI into an infinite number of unconditional moment inequalities using nonnegative weighting functions. Although their test is $\sqrt{n}$-consistent against some alternatives, it follows from [6] that their test has relatively low rate of uniform consistency. The method of [14] is based on estimating moment functions non-parametrically. Their test has good uniform power but implementing their test requires knowledge of certain smoothness properties of moment functions. On the other hand, an advantage of their test is that it becomes very efficient if moment functions are sufficiently smooth (for example, if moment functions are at least twice continuously differentiable).[2,3] I note here that the results in this paper are obtained under similar conditions as those used in [3] and [14], and so the results are different because different testing procedures and different proof techniques are used.

The test of [24] is closely related to that of [3]. [26] 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. [6] developed a new method for computing the critical value for the test statistic of [3] that leads to a more powerful test than theirs but the resulting test is not robust in the sense that it may yield large size distortions in the case of weak instruments. [7], 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. In addition, the critical value in that paper is infeasible since it has the form $a_n\sqrt{\log n/n}$ where a_n is some sequence of unknown positive numbers that is bounded away from zero. After this paper had been made publicly available, [8] made an important contribution by constructing a test based on a statistic

[2]Efficiency of the test of [14] is achieved by using higher order kernel or series methods for estimating moment functions; both the test of [3] and the test developed in this paper work with positive kernels, which exclude higher order kernels, and do not have this feature of the test of [14].

[3]In the statistics literature, there recently have been developed techniques for adaptively selecting the appropriate smoothing parameter for tests like that in [14]. An example is Lepski's method combined with the sample splitting where a part of the sample is used to select the smoothing parameter according to the Lepski's algorithm and the other part is used for testing; see, for example, [18]. Deriving formal results on how the test of [14] works in combination with these adaptive smoothing parameter selection techniques would be an important direction for future research.

that is closely related to that used in this paper with the critical value derived from the limit distribution. The results of that paper complement the results of this paper because [8] obtained an explicit form of the limit distribution that was previously unknown. However, they only provided a plug-in critical value, and it is not obvious how to extend their methods to derive a moment selection procedure.[4]

Finally, an important related paper in the statistics literature is [16]. 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 previous 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. Examples include analysis of incomplete models of English auctions, estimation with interval data, and inference in treatment effect models.

Incomplete Models of English Auctions. My first example follows [19]'s treatment of English auctions under weak conditions. The popular model of English auctions suggested by [29] 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. [19]

[4]Also, it should be noted that [8] imposed rather strong assumptions. Specifically, they assumed existence of finite moment generating function of the noise. In contrast, I only assume 4 finite moments, which is much more plausible in applications. The optimality properties of their test are similar to those of the test developed in this paper.

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 distribution $F(\cdot, X)$ where X denotes observable characteristics of the object. Let $b_1, ..., b_m$ denote highest bids of each bidder. Let $b_{1:m} \leqslant ... \leqslant b_{m:m}$ denote the ordered sequence of bids $b_1, ..., b_m$. Assuming that bids do not exceed bidders' valuations, [19] derived the following upper bound on $F(\cdot, X)$:

$$E[\phi^{-1}(F(v, X)) - I\{b_{i:m} \leqslant v\}|X] \leqslant 0 \, a.s. \tag{2}$$

for all $v \in \mathbb{R}$ and $i = 1, ..., m$ where $\phi(\cdot)$ is a certain (known) increasing function, see equation (3) in [19]. 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 considering (2) for a finite set $\mathcal{V} = \{v_1, ..., v_p\}$ of values of v gives inequalities of the form (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

$$Y = f(X, V) + \varepsilon$$

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 [28], 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

$$E[I\{V_1 \leqslant v\}(Y - f(X, v))|X, V_0, V_1] \leqslant 0 \tag{3}$$

and

$$E[I\{V_0 \geqslant v\}(Y - f(X, v))|X, V_0, V_1] \geqslant 0 \tag{4}$$

for all $v \in \mathbb{R}$. Again, parameterizing the function $f(\cdot, \cdot)$ and selecting a finite set $\mathcal{V} = \{v_1, ..., v_p\}$ gives inequalities of the form (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 counter-factual 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 characteristics X. In other words, the null hypothesis, H_0, is that

$$E[Y_1 - Y_0|X] \geqslant 0 \, a.s. \tag{5}$$

Since in randomized experiments D is independent of X, [26] showed that

$$E[Y_1 - Y_0|X] = E[DY/p - (1-D)Y/(1-p)|X]. \tag{6}$$

Combining (5) and (6) gives CMI of the form (1).

3. 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,

$$f_j(X) \leqslant 0 \text{ for all } j = 1, ..., p \text{ } a.s$$

where $f_j(\cdot)$ denotes the jth component of $f(\cdot)$. 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. $(i = 1, ..., n)$.

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 $\widehat{\Sigma}_i$ of Σ_i. The feasible version of the test will be based on substituting $\widehat{\Sigma}_i$ 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

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

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

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

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

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

where Σ_{j,m_1m_2} denotes the (m_1, m_2) component of Σ_j.

Next, consider a finite set of bandwidth values $H = \{h = h_{\max}a^k : h \geqslant h_{\min}, k = 0, 1, 2, ...\}$ 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 bounded away from zero. For practical purposes, I recommend setting $K(x) = 0.75(1 - \|x\|^2)$ for $\|x\| \leqslant 1$ and 0 otherwise, $h_{\max} = \max_{i,j=1,...,n} \|X_i - X_j\|/2$, $h_{\min} = 0.2h_{\max}(\log n/n)^{1/(3d)}$, and $a = 0.5$.[6] 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 roughly 15 data points when $n = 250$.

Denote $S = \{(i, m, h) : i = 1, ..., n, m = 1, ..., p, h \in H\}$. Based on this notation, the test statistic is

$$T = \max_{s\in S} \frac{\widehat{f}_s}{V_s}.$$

Thus, the test statistic is based on the studentized kernel estimates of the function f at points $\{X_i : i = 1, ..., n\}$.[7]

Let me now explain why the optimal bandwidth value depends on the smoothness properties of the components $f_1, ..., f_p$ of f. Without loss of generality, consider f_1. Suppose that $f_1(X)$ is nearly flat in the neighborhood of its maximum. Then $f_1(X)$ is positive on a large subset of its domain whenever its maximal value is positive. Hence, the maximum of 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 T, and small bandwidth values should be used. I circumvent the problem of bandwidth selection by considering many different 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 many possible degrees of smoothness for $f(X)$.

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

[6]The size of the test is controlled well for many different values of parameter a.

[7]In principle, to form a test statistic, one could specify another grid of points at which the function f would be estimated instead of $\{X_i : i = 1, ..., n\}$. I find it convenient, however, to use $\{X_i : i = 1, ..., n\}$ because this set naturally covers the support of X.

An important feature of the test is that it uses *local* optimal bandwidth values, that is, for each i and m, the test looks for an optimal bandwidth value separately. This is in contrast with *global* optimal bandwidth value that could come, for example, from cross-validation. The use of local optimal bandwidth values is important when the smoothness of the functions $f_m(x)$ is varying over m and x.

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

$$\widehat{T} = \max_{s \in S} \frac{\widehat{f}_s}{\widehat{V}_s}$$

instead of T, where $\widehat{\Sigma}_j$ 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 $\widehat{T}$ 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 $\widehat{T} > c_{1-\alpha}$.

The first method relies on three observations. First, one can approximate $\widehat{T}$ by T. Second, 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 T over all possible functions f satisfying $f \leqslant 0_p$ a.s. corresponds to $f = 0_p$. Third, Lemma 11 in the Appendix shows that the distribution of T is asymptotically independent of the distribution 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 $\tilde{Y}_i \sim N(0_p, \widehat{\Sigma}_i)$ independently across i.
(2) Calculate $T^{PIA} = \max_{(i,m,h) \in S} \sum_{j=1}^n w_h(X_i, X_j)\tilde{Y}_{j,m}/\widehat{V}_{(i,m,h)}$.
(3) Repeat steps 1 and 2 independently B times for some large B to obtain $\{T^{PIA}_b : b = 1, ..., B\}$.
(4) Let $c^{PIA}_{1-\alpha}$ be $(1-\alpha)$ empirical quantile of $\{T^{PIA}_b\}_{b=1}^B$.

The second method is based on the refined moment selection (RMS) procedure. It gives a more powerful test and still controls the required size. The method relies on the observation that $|\widehat{T}| = O_p(\sqrt{\log n})$ if $f = 0_p$ (see Lemmas 8, 9, and 11 in the Appendix) while $\widehat{f}_{i,m,h}/\widehat{V}_{(i,m,h)} \to -\infty$ 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 $\widehat{T}$, so we can ignore corresponding terms in the simulated statistic. Therefore, one can simulate $c_{1-\alpha}$ (denoted by $c^{RMS}_{1-\alpha}$) 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

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

Third, run the following procedure:

(1) For each $i = 1, ..., n$, simulate $\tilde{Y}_i \sim N(0_p, \widehat{\Sigma}_i)$ independently across i.
(2) Calculate $T^{RMS} = \max_{(i,m,h)\in S^{RMS}} \sum_{j=1}^n w_h(X_i, X_j)\tilde{Y}_{j,m}/\widehat{V}_{(i,m,h)}$.
(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 $(1-\alpha)$ empirical quantile of $\{T_b^{RMS}\}_{b=1}^B$.

In the next section, it will be assumed that $\gamma = \gamma_n \to 0$ as $n \to \infty$. 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 [14].[8]

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 [32], [30], [21], and [17]. For scalar-valued Y_i, available estimators are described in [22]. All those estimators can be immediately generalized to vector-valued Y_i's. For concreteness, I describe one estimator here. Choose a bandwidth value $b_n > 0$. For $i = 1, ..., n$, let $J(i) = \{j = 1, ..., n : \|X_j - X_i\| \leqslant b_n\}$. 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 number of elements in $J(i)$. Then Σ_i can be estimated by

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

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.

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

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

[8]Note also that if γ is comparable with α, it follows from the proof of Theorem 1 below that one can do a finite sample adjustment of the critical value by taking $(1 - \alpha + 2\gamma)$ quantile of $\{T_b^{RMS}\}_{b=1}^B$ at step 4 of the procedure above. If this is done, then the result (7) in Theorem 1 below holds without Assumption A7 stated in the next section, so that one does not have to assume that $\gamma = \gamma_n \to 0$. Also, the theory in the next section requires that $\gamma \leqslant Cn^{-c}$ for some constants c and C. Nevertheless, in my Monte Carlo simulations I use the rule $\gamma = 0.1/\log(n)$ to make meaningful comparisons with [14].

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,

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

so that

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

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 $\|X_{k(j)} - X_i\| \leqslant b_n$ and $\|X_j - X_i\| \leqslant b_n$.

4. Main Results

This section presents 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. 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 and C_j for $j = 1, ..., 5$ be strictly positive and finite constants independent of the sample size n. Let $M_h(X_i)$ be the number of elements in the set $\{X_j : \|X_j - X_i\| \leqslant h,\ j = 1, ..., n\}$. Results in this paper will be proven under the following assumptions.

A 1. *(i) Design points $\{X_i\}_{i=1}^n$ are non-stochastic. (ii) $c_1 n h^d \leqslant M_h(X_i) \leqslant C_1 n h^d$ for all $i = 1, ..., n$ and $h \in H = H_n$.*

The design points are non-stochastic because the analysis is conducted conditional on X_i's. In addition, A1 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 [22] that A1 holds in an i.i.d. setting with probability approaching one as the sample size increases if the distribution of X_i is absolutely continuous with respect to Lebesgue measure, has bounded support, and has density bounded away from zero on the support. This statement is not precise unless one makes some extra assumptions. Lemma 3 in the Appendix gives a counter-example. Instead, Lemma 4 shows that A1 holds for large n a.s. if, in addition, it is assumed 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 A1 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".

A 2. *(i) Disturbances* $\{\varepsilon_i\}_{i=1}^n$ *are independent* $\mathbb{R}^p$*-valued random vectors with* $E[\varepsilon_i] = 0$ *for all* $i = 1, ..., n$*. (ii)* $E[\max_{m=1,...,p} |\varepsilon_{i,m}|^4] \leqslant C_2$ *for all* $i = 1, ..., n$*. (iii)* $\Sigma_{i,mm} \geqslant c_2$ *for all* $i = 1, ..., n$ *and* $m = 1, ..., p$*.*

Finite fourth moment of disturbances is used to show that the distribution of the test statistic $\widehat{T}$ in large samples does not depend on the form of the distribution of ε_i's. 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.

Let $\tau, L > 0$ be arbitrary positive numbers. Let $[\tau]$ denote the largest integer strictly smaller than τ. Before stating A3, let me give formal definitions of Holder smoothness classes $\mathcal{F}_0(\tau, L)$ and their subsets $\mathcal{F}(\tau, L)$ to be considered in this paper. 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

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

whenever it exists. It is said that the function $g : \mathbb{R}^d \to \mathbb{R}$ belongs to the class $\mathcal{F}_0(\tau, L)$ if (i) g has continuous partial derivatives up to order $[\tau]$, (ii) for any $\alpha = (\alpha_1, ..., \alpha_d)$ such that $|\alpha| = [\tau]$ and $x, y \in \mathbb{R}^d$,

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

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

$$|D^\alpha g(x)| \leqslant L.$$

The function classes $\mathcal{F}(\tau, L) \subset \mathcal{F}_0(\tau, L)$ are defined as follows. Let $S^{d-1} = \{l \in \mathbb{R}^d : \|l\| = 1\}$ denote the space of directions in $\mathbb{R}^d$. For any $g \in \mathcal{F}_0(\tau, L)$, $x = (x_1, ..., x_d) \in \mathbb{R}^d$, and $l \in S^{d-1}$, let $g^{(k,l)}(x)$ denote kth derivative of function g in direction l at point x whenever it exists.[9] Then $\mathcal{F}(\tau, L)$ consists of all elements g of $\mathcal{F}_0(\tau, L)$ satisfying $g^{(k,l)}(x) = 0$ for all $k = 1, ..., [\tau]$ whenever $g^{(1,l)}(x) = 0$. Note that for $\tau \leqslant 2$, $\mathcal{F}(\tau, L) = \mathcal{F}_0(\tau, L)$.

A3. *Components* f_m*'s of the regression function* f *satisfy* $f_m \in \mathcal{F}(\tau, L)$ *for all* $m = 1, ..., p$*.*

For simplicity of notation, I assume that all components of f have the same smoothness properties. This assumption is only used in the derivation of the power properties of the test. The reason I consider subsets $\mathcal{F}(\tau, L)$ of $\mathcal{F}_0(\tau, L)$ is related to the fact that I only consider positive kernels; see A5 below. Positive kernels exclude higher-order kernels, which have to be used to effectively reduce bias of estimating sufficiently smooth nonparametric functions. Therefore, for

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

$\tau > 2$, I have to exclude certain functions from the function class $\mathcal{F}_0(\tau, L)$ to achieve minimax rate optimality.

A 4. *(i) The set of bandwidth values has the following form:* $H = H_n = \{h = h_{\max} a^k : h \geqslant h_{\min}, k = 0, 1, 2, ...\}$ *where* $a \in (0, 1)$, $h_{\max} = \max_{i,j=1,...,n} \|X_i - X_j\|/2$ *and* $h_{\min} = C_3(\log n/n)^{1/(3d)}$. *(ii)* $S = S_n = \{(i, m, h) : i = 1, ..., n, m = 1, ..., p, h \in H_n\}$.

According to this assumption, the maximal bandwidth value, $h_{\max}$, is chosen to match the radius of the support of design points. It is intended 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 at an appropriate rate. The minimal bandwidth value is intended to detect alternatives with narrow peaks. A4 is a key condition used to establish an invariance principle that shows that the distribution of $\widehat{T}$ asymptotically depends on the distribution of disturbances ε_i's only through their covariances Σ_i's. Note also that the choice of $h_{\max}$ in this assumption is made only for convenience. The result of Theorem 1 below would apply in the case when $h_{\max} \to 0$ as well. However, if $h_{\max} \to 0$, then the test will have relatively low power in detecting global deviations from H_0 where one of the functions $f_1, ..., f_p$ is positive on a large set.

A 5. *(i) The kernel* K *is positive and supported on* $\{x \in \mathbb{R}^d : \|x\| \leqslant 1\}$. *(ii)* $K(x) \leqslant 1$ *for all* $x \in \mathbb{R}^d$ *and* $K(x) \geqslant c_3$ *for all* $\|x\| \leqslant 1/2$.

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 [36] 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 the minimax optimal testing rate over classes $\mathcal{F}_0(\tau, L)$. 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 under-smoothing so that the bias of the estimator is asymptotically negligible in comparison with its standard deviation. Otherwise, large values of $\widehat{T}$ 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 under-smoothing, one has to know the smoothness properties of $f(X)$. In contrast, 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 bounded away from zero. Nevertheless, the test developed in this paper will be rate optimal in the minimax sense against classes $\mathcal{F}(\tau, L)$ when $\tau > d$. In addition, I will explain that if A2 is strengthened, so that ε_i's are sub-exponential, then one can set $h_{\min} \asymp n^{-d}$ (up-to logs) in A4, and the test will be rate optimal in the minimax sense against classes $\mathcal{F}(\tau, L)$ with $\tau \leqslant d$ as well.

A6. *Estimators* $\widehat{\Sigma}_i$ *of* Σ_i *satisfy* $\mathrm{P}(\max_{i=1,...,n} \|\widehat{\Sigma}_i - \Sigma_i\| > C_4 n^{-c_4}) \leqslant C_4 n^{-c_4}$ *where* $\|\cdot\|$ *denotes the spectral norm on the space of* $p \times p$*-dimensional symmetric matrices corresponding to the Euclidean norm on* $\mathbb{R}^p$*.*

A6 is satisfied for $\widehat{\Sigma}_i$ described in Section 3.3; see Lemma 1 in the Appendix. In practice, due to the curse of dimensionality, it might be useful to use some parametric or semi-parametric estimators of Σ_i's 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 [32] (or its multivariate generalization) is $1/\sqrt{n}$-consistent. In this case, A6 will be satisfied with $\kappa = 1/4 - \phi$ for arbitrarily small $\phi > 0$.

A7. *The truncation parameter* γ *satisfies* $\gamma = \gamma_n \leqslant C_5 n^{-c_5}$*.*

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

A1-A3 concern the data-generating process while A4-A7 deal with the test. Taken all together, they define the model.[10] The asymptotic results in this paper will be shown to hold uniformly over all models satisfying A1-A7. For that purpose, the following notation will be useful. Let $\mathcal{G}$ denote the set of all models satisfying A1-A7 for all n, and let $w \in \mathcal{G}$ denote a generic model in $\mathcal{G}$. In addition, let $E_w[\cdot]$ denote the expectation calculated assuming the model w. Finally, let $f(w)$ denote the regression function f corresponding to the model 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 limit distribution. Instead, I use a finite sample method developed in [13]. For each sample size n, this method gives an upper error bound on the uniform distance between the cdf of the test statistic and the cdf the test statistic would have in the model with Gaussian noise $\{\varepsilon_i\}_{i=1}^n$.

Let $\mathcal{G}_0$ and $\mathcal{G}_{00}$ denote the set of all elements w of $\mathcal{G}$ satisfying $f(w) \leqslant 0$ a.s. and $f(w) = 0$ 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* $P = PIA$ *or* RMS*. Then for some constants* c *and* C *depending only on* c_j *and* C_j *for* $j = 1, ..., 5$*,*

$$\sup_{w \in \mathcal{G}_0} P_w \left(\widehat{T} > c^P_{1-\alpha} \right) \leqslant \alpha + Cn^{-c} \text{ for all } n. \tag{7}$$

[10]The model in this paper is understood as infinite sequences of non-stochastic design points $\{X_i\}_{i=1}^\infty$ and random disturbances $\{\varepsilon_i\}_{i=1}^\infty$, a vector-valued regression function f, sequence of estimators $\{\widehat{\Sigma}_i\}_{i=1}^n$ for each n, a kernel K, a sequence of sets of bandwidth values $\{H_n\}_{n=1}^\infty$, a sequence of sets $\{S_n\}_{n=1}^\infty$, and a sequence of truncation parameters $\{\gamma_n\}_{n=1}^\infty$.

In addition,

$$\inf_{w \in \mathcal{G}_{00}} P_w \left(\widehat{T} > c^P_{1-\alpha} \right) \geqslant \alpha - Cn^{-c} \text{ for all } n. \tag{8}$$

Comment 1. (i) Proofs of all results are presented in the Appendix.

(ii) An advantage of this theorem is that it shows that the probability of rejecting the null under the null can exceed the nominal level of the test only by a *polynomially small* (in n) number. This implies that the bootstrap procedures developed in this paper provide high quality inference in finite samples. All other papers on CMI only provide results that the probability of rejecting the null under the null is *asymptotically* not larger than the nominal level, without providing a bound on the difference between two quantities.

(iii) The theorem provides a bound on the difference in the probability of rejecting the null and the nominal level that holds *uniformly* over a large class of models. This also serves as a guarantee that the test controls size well in finite samples.

(iv) Combining (7) and (8) shows that uniformly over $w \in \mathcal{G}_{00}$, $|P_w(\widehat{T} > c^P_{1-\alpha}) - \alpha| \leqslant Cn^{-c}$.

4.3. **Consistency Against a Fixed Alternative.** Consider any model $w \in \mathcal{G}$. Let $f = f(w)$. I will consider the following distance between the model w and the null hypothesis:

$$\rho(w, H_0) = \sup_{i=1,...,\infty;\, m=1,...,p} [f_m(X_i)]_+ \tag{9}$$

For any alternative outside of the set Θ_I, $\rho(w, H_0) > 0$. In the definition of $\rho(w, H_0)$, I take the supremum over all $i = 1, ..., \infty$ because I consider non-stochastic X_i's. Alternatively, if X_i's were sampled from some distribution with the support $\mathcal{X}$, I would set $\rho(w, H_0) = \sup_{x \in \mathcal{X}; m=1,...,p} [f_m(x)]_+$. The following theorem shows that the test is consistent against any fixed alternative $w \in \mathcal{G}$ with $\rho(w, H_0) > 0$. Moreover, the theorem shows 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 $\rho(w, H_0) \geqslant \rho$.

Theorem 2. *Let $P = PIA$ or RMS. Then for some constants c and C depending only on c_j and C_j for $j = 1, ..., 5$,*

$$\inf_{w \in \mathcal{G}_\rho} P_w \left(\widehat{T} > c^P_{1-\alpha} \right) \geqslant 1 - Cn^{-c} \text{ for all } n. \tag{10}$$

4.4. **Consistency Against Local One-Dimensional Alternatives.** This section derives the rate of consistency of the test against one-dimensional alternatives. Consider any model $w_0 \in \mathcal{G}$ such that $\rho(w_0, H_0) > 0$. For some sequence $\{a_n\}_{n=1}^\infty$ of positive numbers converging to zero, consider the sequence of models $\{w_n\}_{n=1}^\infty$ such that for all n, w_n coincides with w_0 except that $f(w_n) = a_n f(w_0)$. I refer to such sequences as local one-dimensional alternatives. The following theorem establishes the consistency of the test against such alternatives whenever $a_n \sqrt{n/\log n} \to \infty$.

Theorem 3. *Let $P = PIA$ or RMS. Assume that $a_n\sqrt{n/\log n} \to \infty$. Then*

$$P_{w_n}\left(\widehat{T} > c^P_{1-\alpha}\right) \to 1 \text{ as } n \to \infty. \tag{11}$$

Comment 2. The test of [3] is consistent against the same sequence of alternatives if $a_n\sqrt{n} \to \infty$. Hence, my test is consistent against nearly the same sequence of one-dimensional alternatives as the test of [3]. 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 $\mathcal{F}(\tau, L)$ under certain additional constraints. These additional constraints are needed to deal with 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. $B_\vartheta(x) \subset S$. When $\tau \leqslant d$, set $\vartheta = \vartheta_n = 2\sqrt{d}h_{\min}$. When $\tau > d$, set $\vartheta = \vartheta_n = 2\sqrt{d}(\log n/n)^{1/(2\tau+d)}$. Let $\mathbb{N}_\vartheta = \{i \in \mathbb{N} : X_i \in S_\vartheta\}$. For any $w \in \mathcal{G}$ and corresponding $f = f(w)$, let

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

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, [16] and [26]. Let $\mathcal{G}_\vartheta$ be the subset of all elements of $\mathcal{G}$ such that $\inf_{w \in \mathcal{G}_\vartheta} \rho_\vartheta(w, H_0)/h^\tau_{\min} \to \infty$ if $\tau \leqslant d$ and $\inf_{w \in \mathcal{G}_\vartheta} \rho_\vartheta(w, H_0)(n/\log n)^{\tau/(2\tau+d)} \to \infty$ if $\tau > d$. Then

Theorem 4. *Let $P = PIA$ or RMS. Then*

$$\inf_{w \in \mathcal{G}_\vartheta} P_w\left(\widehat{T} > c^P_{1-\alpha}\right) \to 1 \text{ as } n \to \infty. \tag{12}$$

Comment 3. Consider a model with $m(X, W, \theta) = \tilde{m}(X, W) + \theta$. 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 $\Theta_I = \{\theta \in \mathbb{R} : \theta \leqslant 0\}$. The theorem above shows that the test developed in this paper is consistent against sequences of alternatives $\theta_0 = \theta_{0,n}$ whenever $\theta_{0,n}(n/\log n)^{\nu/(2\nu+1)} \to \infty$. At the same time, it follows from [6] that the test of [3] is consistent only if $\theta_{n,0}n^{\nu/(2\nu+2)} \to \infty$, so their test has a slower rate of consistency than that developed in this paper against this sequence of alternatives 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 A1, let $\mathcal{G}_X$ denote the set of all models w in $\mathcal{G}$ with the sequence of design points X. 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 $\|x_i - x_j\| \geqslant h$

for all $i, j = 1, ..., m$ if $i \neq j$. I will assume that $N(h, S_\vartheta) \geqslant Ch^{-d}$ 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 $\phi_n(Y_1, ..., Y_n)$, $n \geqslant 1$, denote a sequence of tests. In other words, $\phi_n(Y_1, ..., Y_n)$ denotes the probability of rejecting the null hypothesis upon observing sample $Y = (Y_1, ..., Y_n)$.

Theorem 5. *Assume that (i) $N(h, S_\vartheta) \geqslant Ch^{-d}$ for all $h \in (0, 1)$, sufficiently large n, and some $C > 0$, and (ii) $r_n(n/\log n)^{\tau/(2\tau+d)} \to 0$ 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 $\sup_{w \in \mathcal{G}_0 \cap \mathcal{G}_X} E_w[\phi_n(Y_1, ..., Y_n)] \leqslant \alpha + o(1)$,*

$$\inf_{w \in \mathcal{G}_X, \rho_\vartheta(w, H_0) \geqslant Cr_n} E_w[\phi_n(Y_1, ..., Y_n)] \leqslant \alpha + o(1) \text{ as } n \to \infty. \tag{13}$$

Comment 4. Since $\mathcal{G}_X \subset \mathcal{G}$, the same lower bound also applies with $\mathcal{G}$ instead of $\mathcal{G}_X$ as well. Comparing this result with that in Theorem 4 shows that the test presented in this paper is minimax rate optimal in the class $\mathcal{F}(\tau, L)$ (for almost all sequences $\{X_i\}_{i=1}^\infty$) if $\tau > d$. The lower bound is not achieved when $\tau \leqslant d$. In this case, the test does not achieve the lower bound because of the constraint on $h_{\min}$ imposed in A4. It is unknown whether this constraint can be relaxed without imposing existence of higher moments of ε_i's beyond those imposed in A2(ii). However, under the assumption of finite moment generating function of ε_i's, the test developed in this paper achieves the optimal rate. Specifically, assume in A2 that $E[\exp(|\varepsilon_{i,m}|/C_2)] \leqslant C_2$ for all $i = 1, ..., n$ and $m = 1, ..., p$ instead of $E[\max_{m=1,...,p} |\varepsilon_{i,m}^4|] \leqslant C_2$ for all $i = 1, ..., n$ and assume in A4 that $h_{\min} = C_3(\log n)^8/n^d$ instead of $h_{\min} = C_3(\log n/n)^{1/(3d)}$. Then Theorem 1 continues to hold (with the only difference in the proof that now one applies Corollary 2.3, case E.3 in [13] instead of case E.4 in Lemma 11) and the argument like that in the proof of Theorem 4 shows that the test is uniformly consistent against the set of models $\mathcal{G}_\vartheta$ if $\inf_{w \in \mathcal{G}_\vartheta} \rho_\vartheta(w, H_0)(n/\log n)^{\tau/(2\tau+d)} \to \infty$. This implies that the test is rate optimal when $\tau \leqslant d$ as well.

5. Extensions

In this section, I briefly outline two extensions 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) and (3)-(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 numbers. Suppose that $\theta \in \Theta$ satisfies

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

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 $\{p_n\}_{n=1}^{\infty}$ be a sequence of natural numbers converging to infinity. Consider the test based on the first $p = p_n$ inequalities. More precisely, let $m : \mathbb{R}^d \times \mathbb{R}^k \times \Theta \to \mathbb{R}^p$ be the vector-valued function whose jth component coincides with jth component of $\tilde{m}$ for all $j = 1, ..., p$, and consider the test described in Section 3 based on inequalities $E[m(X, W, \theta)|X] \leqslant 0$ a.s. Denote its critical value by $c_{1-\alpha}^P$ with $P = PIA$ or RMS. In addition, let me use all the notation defined in Sections 3 and 4 and corresponding to the test based on the function m.

Let $\mathcal{G}$ denote the set of all models satisfying A1-A7 for all n where $p = p_n$ is now understood to be a function of n and where ε_i, Σ_i, f, and p in A2 and A3 are replaced by $\tilde{\varepsilon}_i$, $\tilde{\Sigma}_i$, $\tilde{f}$, and ∞, respectively. Let $\mathcal{G}_0$ denote the set of models in $\mathcal{G}$ satisfying $\tilde{f} \leqslant 0$ a.s., $\mathcal{G}_{00}$ denote the set of models in $\mathcal{G}_0$ satisfying $\tilde{f} = 0$ a.s., and $\mathcal{G}_\rho$ denote the set of models in $\mathcal{G}$ satisfying $\rho(w, H_0) \geqslant \rho$ where $\rho(w, H_0)$ is defined as in (9) with $\tilde{f}$ and ∞ instead of f and p, respectively. 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 $P = PIA$ or RMS. Assume that $p_n \log n \leqslant C_6 n^{c_6}$ for some sufficiently small and large constants c_6 and C_6, respectively.*[11]*. Then for some constants c and C depending only on c_j and C_j for $j = 1, ..., 6$,*

$$\sup_{w \in \mathcal{G}_0} P_w \left(\widehat{T} > c_{1-\alpha}^P \right) \leqslant \alpha + Cn^{-c} \text{ for all } n. \tag{14}$$

In addition,

$$\inf_{w \in \mathcal{G}_{00}} P_w \left(\widehat{T} > c_{1-\alpha}^P \right) \geqslant \alpha - Cn^{-c} \text{ for all } n. \tag{15}$$

Finally,

$$\inf_{w \in \mathcal{G}_\rho} P_w \left(\widehat{T} > c_{1-\alpha}^P \right) \geqslant 1 - Cn^{-c} \text{ for all } n. \tag{16}$$

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

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

$$E[m(X, W, \theta)|X^1, X^2 = x_0] \leqslant 0 \, a.s. \tag{17}$$

where $m(\cdot, \cdot, \cdot)$, $X = (X^1, X^2)$, and W are as above, and x_0 is some fixed point of interest. Assume that X^1 and X^2 are d_1- and d_2-dimensional random vectors (the dimension of X is

[11] Inspection of the proof shows that it suffices to choose $c_6 < c_4/8$ and some $C_6 > 0$.

$d = d_1 + d_2$). CMI of the form (17) arise in nonparametric and semi-parametric inference. For example, recall the English auction model from Section 2. In that model, suppose that the set of covariates is $X = (X^1, X^2)$ so that $F = F(v, X^1, X^2)$. Suppose that the point $X^2 = x_0$ is of interest. Denote $\tilde{F}(v, X^1) = F(v, X^1, x_0)$. Then inequality (2) leads to

$$E[\phi^{-1}(\tilde{F}(v, X^1)) - I\{b_{i:m} \leqslant v\}|X^1, X^2 = x_0] \leqslant 0 \, a.s.$$

Parameterizing the function $\tilde{F}(\cdot, \cdot)$ gives inequalities of the form (17). 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 X^2 is x_0.

Models based on local CMI were previously studied in [14] and in [4]. The approach of [4] is based on the results of [3], and so is different from the approach taken here. The approach of [14] is similar to their techniques used for testing CMI (1), hence the same comments on their approach apply here as those given in the Introduction.

As above, given θ_0, define $f(X) = E[m(X, W, \theta_0)|X]$. In addition, denote $\varepsilon_i = m(X_i, W_i, \theta_0) - f(X_i)$, and $\Sigma_i = E[\varepsilon_i(\varepsilon_i)^T|X_i]$. Let $\widehat{\Sigma}_i$ be an estimator of Σ_i $(i = 1, ..., n)$ as described in Section 3.3. Let N be a subset of all observations $i = 1, ..., n$ such that $\|X_i^2 - x_0\| < a$ 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 $i = 1, ..., n_a$. In order to test inequalities (17), consider the test described in Section 3 based on the data $\{(X_i, W_i)\}_{i \in N}$. Denote its test statistic by $\widehat{T}$ and its critical value by $c_{1-\alpha}^P$ with $P = PIA$ or RMS.

Let $\mathcal{G}$ denote the set of models satisfying A1-A7 for all n with n replaced by n_a, with d replaced by d_1 in A1 and A4, and such that for all these models, $|f_m(X)| \leqslant C_6 a_n$ for all $i \in N$ and $m = 1, ..., p$, and $a_n \sqrt{n_a h_{\max}^{d_1} \log n} \leqslant C_6 n^{-c_6}$ for sufficiently small and large constants c_6 and C_6, respectively. Let $\mathcal{G}_0$ denote the set of all models in $\mathcal{G}$ satisfying $f(X) \leqslant 0$ a.s., $\mathcal{G}_{00}$ denote the set of models in $\mathcal{G}_0$ satisfying $f(X) = 0$ a.s. Denote $\mathcal{N}_a = \{(i, m) : i = 1, ..., \infty, m = 1, ..., p, \|X_i^2 - x_0\| \leqslant a\}$. Define the distance between the model $w \in \mathcal{G}$ and the null hypothesis by

$$\rho(w, H_0) = \inf_{a \in (0, \infty)} \sup_{(i,m) \in \mathcal{N}_a} [f_m(X_i)]_+$$

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

Corollary 2. *Let $P = PIA$ or RMS. Then for some constants c and C depending only on c_j and C_j for $j = 1, ..., 6$,*

$$\sup_{w \in \mathcal{G}_0} P_w \left(\widehat{T} > c_{1-\alpha}^P \right) \leqslant \alpha + Cn^{-c} \text{ for all } n. \tag{18}$$

In addition,

$$\inf_{w \in \mathcal{G}_{00}} P_w \left(\widehat{T} > c_{1-\alpha}^P \right) \geqslant \alpha - Cn^{-c} \text{ for all } n. \tag{19}$$

Finally,

$$\inf_{w\in\mathcal{G}_\rho} P_w\left(\widehat{T} > c^P_{1-\alpha}\right) \geqslant 1 - Cn^{-c} \text{ for all } n. \tag{20}$$

Comment 6. (i) Note that in an i.i.d. setting, if $f(x^1, x_0) > 0$ for some x^1 such that (x^1, x_0) is inside of the support of X, then it follows as in the proof of Lemma 4 that $\rho(w, H_0) > 0$ 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 $a_n\sqrt{n_a h_{\max}^{d_1}\log n} \leqslant C_6 n^{-c_6}$ in this corollary is required to ensure that the bias due to using data with $X_i^2 \neq x_0$ is asymptotically negligible. Given that small values of a_n 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 [3], [14], and [26]. The methods of [3] and [26] 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 [14] has higher power against latter 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 [3] and [26], my test has higher power against alternatives with peaks. In comparison with [14], my test has higher power when their rule-of-thumb techniques lead to an inappropriate bandwidth value.[12] 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 [27]).

First simulation study. The data generating process is

$$Y_i = L(M - |X_i|)_+ - m + \varepsilon_i$$

where X_i's are equidistant on the $[-2, +2]$ interval[13], Y_i's and ε_i's are scalar random variables, and L, M, and m are some constants. Depending on the experiment, ε_i's have either normal or (continuous) uniform distribution with mean zero. In both cases, the variance of ε_i's is 0.01. I consider the following specifications for parameters. Case 1: $L = M = m = 0$. Case 2: $L = 0.1$,

[12]When their rule-of-thumb works well and moment functions are sufficiently smooth, the test of [14] often yielded the best results in my simulations.

[13]Results where X_i's are distributed uniformly on the $[-2, +2]$ interval are very similar.

$M = 0.2$, $m = 0.02$. Case 3: $L = M = 0$, $m = -0.02$. Case 4: $L = 2$, $M = 0.2$, $m = 0.2$. Note that $E[Y|X] \leqslant 0$ 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. 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 [3], 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 asymptotic approximation suggested in their paper. 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[14]

$$K(x) = 1.5(1 - 4x^2)_+.$$

For the test of [14], I use their kernel type test statistic with critical values based on the multiplier bootstrap both with (CLR, $\widehat{V}$) and without (CLR, V) the set estimation. Both [14] and [26] (LSW) circumvent edge effects of kernel estimators by restricting their test statistics to the proper subsets of the support of X. To accommodate 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 [26], I use their test statistic based on one-sided L_1-norm.

Parameters for the test developed in this paper are chosen according to recommendations in Section 3.1. Specifically, the largest bandwidth value, $h_{\max}$, is set to be equal to the length of the support of the empirical distribution.[15] The smallest bandwidth value, $h_{\min}$, is set as $h_{\min} = 0.2h_{\max}(\log n/n)^{1/3}$. The scaling parameter, a, equals 0.5 so that the set of bandwidth values is

$$H_n = \{h = h_{\max}0.5^k : h \geqslant h_{\min}, k = 0, 1, 2, ...\}.$$

I estimate Σ_i using the method of [32]. Specifically, I rearrange the data so that $X_1 \leqslant ... \leqslant X_n$ and set $\widehat{\Sigma}_i = \widehat{\Sigma} = \sum_{i=2}^n (Y_i - Y_{i-1})^2/(2n)$. Finally, for the RMS critical value, I set $\gamma = 0.1/\log(n)$

[14]This kernel function does not coincide with recommendations in Section 3.1 (where I recommended the kernel $K(x) = 0.75(1-x^2)_+$). I use this kernel function because it was used in other simulation studies; see, in particular [26]. Note, however, that for the test statistic in this paper, a multiplicative constant in the kernel function has no effect (it cancels out because of studentization), and so using kernels $K(x) = 1.5(1-4x^2)_+$ and $K(x) = 0.75(1-x^2)_+$ gives numerically the same values of the test statistic if all bandwidth values for the former kernel are twice as large as bandwidth values for the latter kernel.

[15]In Section 3.1, I recommend setting the largest bandwidth value as one half of the length of the support of the empirical distribution. The difference is explained by different scaling of the kernel function.

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, $\widehat{V}$	Adaptive test, plugin	Adaptive test, RMS
Normal	1	0.096	0.908	0.108	0.144	0.144	0.100	0.100
	2	0.002	0.005	0.000	0.010	0.010	0.005	0.005
	3	0.880	0.880	0.922	0.803	0.803	0.756	0.756
	4	0.000	0.023	0.000	0.053	0.138	0.803	0.882
Uniform	1	0.102	0.103	0.112	0.142	0.142	0.105	0.124
	2	0.004	0.007	0.001	0.013	0.013	0.003	0.003
	3	0.893	0.893	0.924	0.780	0.780	0.771	0.771
	4	0.000	0.023	0.000	0.038	0.115	0.797	0.867

to make meaningful comparisons with the test of [14]. In all bootstrap procedures, for all tests, I use 500 repetitions.

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 uniform 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 [3] and [26] 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 [14]. 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 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 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 (2013) test, which is most closely related to my method. For my test, I use the RMS critical value. For the test of [3], I use their GMS critical value. The data generating process is

$$Y_i = m + \sqrt{2\pi}\phi(\tau X_i) + \varepsilon_i$$

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, $\widehat{V}$	Adaptive test, plugin	Adaptive test, RMS
Normal	1	0.089	0.091	0.134	0.146	0.146	0.108	0.108
	2	0.001	0.002	0.000	0.000	0.000	0.006	0.006
	3	0.990	0.990	0.996	0.940	0.940	.955	0.955
	4	0.002	0.809	0.000	0.500	0.754	0.994	0.999
Uniform	1	0.083	0.089	0.103	0.116	0.116	0.106	0.106
	2	0.000	0.004	0.000	0.002	0.002	0.003	0.003
	3	0.992	0.992	0.995	0.919	0.919	0.958	0.958
	4	0.003	0.818	0.000	0.474	0.750	0.991	1.000

where X_i's are again equidistant on the $[-2, +2]$ interval, Y_i's and ε_i's are scalar random variables, m and τ are some constants, and $\phi(\cdot)$ is the pdf of the standard Gaussian distribution. In this experiment, ε_i's have $N(0, 1)$ distribution. I use samples $(X_i, Y_i)_{i=1}^n$ of size $n = 250$. Both tests are based on the same specifications as in the first simulation study except that now I use 100 repetitions for all simulation 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 $\tau \geqslant 0$. In addition, $E[Y|X] \leqslant 0$ a.s. if $m \leqslant -1$. Therefore, I set $m \geqslant -1$. Figure 1 shows the difference between the rejection probabilities of my test and of the test of [3]. This figure shows that the rejection probability of the test developed in this paper is higher than that of the test of [3] 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

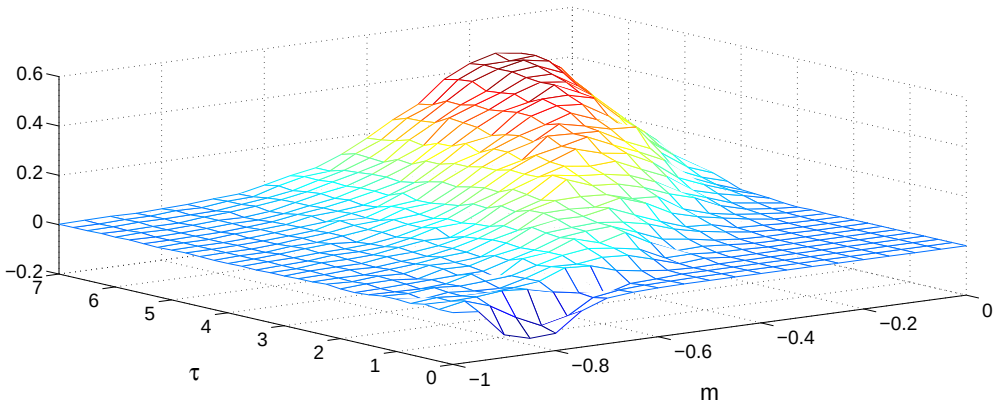

FIGURE 1. The difference between the rejection probabilities of the test developed in this paper and of the test of Andrews and Shi (2013) (with RMS and GMS critical values correspondingly). The nominal size is 10%. Results are based on 500 simulations. The figure shows that the rejection probability of the test developed in this paper is higher than that of the test of Andrews and Shi (2013) in most cases and is strictly higher over a wide region of parameter values.

all, large classes of smooth alternatives whose distance from the null hypothesis converges to zero at a fastest possible rate. The tests of [3] and [26] 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.

APPENDIX A. PROOFS

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 $\widehat{\Sigma}_i$ 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 sufficient conditions for A1 in the main part of the paper. Section A.4 explains an anticoncentration inequality for the maximum of Gaussian random variables with unit variance. Section A.5 describes a result on Gaussian random variables that is used in the proof of the lower bound on the minimax rate. Section A.6 develops some preliminary technical results necessary for the proofs of the main theorems. Finally, Section A.7 presents the proofs of the theorems stated in the main part of the paper.

In this Appendix, c and C are used as generic *strictly* positive constants that are independent of n. Their values can change from line to line.

A.1. **Lemma on the Estimator of Σ_i.**

Lemma 1. *Let $\widehat{\Sigma}_i$ be an estimator of Σ_i described in Section 3.3. Let A1-A3 hold. In addition, assume that (i) $E[|\varepsilon_{i,m}|^{4+\delta}] \leqslant C$ for all $i = 1, ..., n$ and $m = 1, ..., p$, (ii) $b \leqslant Cn^{-c}$, (iii) $\min_{i=1,...,n} |J(i)|/n^{1/(2+\delta)} \geqslant cn^C$, (iv) $\|\Sigma_i - \Sigma_j\| \leqslant C\|X_i - X_j\|$. Then A6 holds.*

Comment 7. Note that under assumptions of Lemma 4, which is described below, condition (iii) above follows from $n^{(1+\delta)/(2+\delta)}b^d \geqslant cn^C$, which is an elementary condition.

Proof. By definition,

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

Since all norms on the finite-dimensional linear space are equivalent (Theorem 1.6 in [25]), it is enough to prove that

$$P\left(\max_{i=1,...,n} |\widehat{\Sigma}_{i,m_1m_2} - \Sigma_{i,m_1m_2}| > Cn^{-c}\right) \leqslant Cn^{-c}$$

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, $\widehat{\Sigma}_i$, $f(X_i)$, and ε_i instead of $\Sigma_{i,11}$, $\widehat{\Sigma}_{i,11}$, $f_1(X_i)$, and $\varepsilon_{i,1}$ correspondingly as if it were a one-dimensional case.

Let $M = n^{1/(4+\delta/2)}$. Consider a truncated version of ε_i's: $\tilde{\varepsilon}_i = \varepsilon_i I\{|\varepsilon_i| \leqslant M\}$. Since $E[|\varepsilon_i|^{4+\delta}] \leqslant C$, it follows that

$$E\left[\max_{i=1,...,n} |\varepsilon_i|\right] \leqslant \left(E\left[\max_{i=1,...,n} |\varepsilon_i|^{4+\delta}\right]\right)^{1/(4+\delta)} \leqslant \left(E\left[\sum_{i=1}^n |\varepsilon_i|^{4+\delta}\right]\right)^{1/(4+\delta)} \leqslant Cn^{1/(4+\delta)}.$$

Then Markov inequality gives

$$P\left(\max_{i=1,...,n} |\varepsilon_i| > M\right) \leqslant Cn^{1/(4+\delta)}/M \leqslant Cn^{-c}.$$

So,

$$\mathcal{P} = P\left(\max_{i=1,...,n} |\tilde{\varepsilon}_i - \varepsilon_i| > 0\right) \leqslant Cn^{-c}.$$

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

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

In addition, for $i = 1, ..., n$, denote $\tilde{Y}_i = f(X_i) + \tilde{\varepsilon}_i$ and

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

Then

$$P\left(\max_{i=1,...,n} |\bar{\Sigma}_i - \widehat{\Sigma}_i| > 0 \right) \leqslant \mathcal{P} \leqslant Cn^{-c}.$$

Therefore, for sufficiently small c and sufficiently large C,

$$P\left(\max_{i=1,...,n} |\widehat{\Sigma}_i - \Sigma_i| > Cn^{-c} \right) \leqslant P\left(\max_{i=1,...,n} |\bar{\Sigma}_i - \tilde{\Sigma}_i| > Cn^{-c}/2 \right) + Cn^{-c}$$

for all n. By the union bound,

$$P\left(\max_{i=1,...,n} |\bar{\Sigma}_i - \tilde{\Sigma}_i| > Cn^{-c} \right) \leqslant \sum_{i=1}^{n} P\left(|\bar{\Sigma}_i - \tilde{\Sigma}_i| > Cn^{-c} \right).$$

Further,

$$P\left(|\bar{\Sigma}_i - \tilde{\Sigma}_i| > Cn^{-c} \right) \leqslant P_1 + P_2 + P_3$$

where

$$P_1 = P\left(\sum_{j \in J(i)} (f(X_{k(j)}) - f(X_j))^2 / (2|J(i)|) > Cn^{-c} \right),$$

$$P_2 = P\left(|\sum_{j \in J(i)} (f(X_{k(j)}) - f(X_j)(\tilde{\varepsilon}_{k(j)} - \tilde{\varepsilon}_j))| / |J(i)| > Cn^{-c} \right),$$

$$P_3 = P\left(|\sum_{j \in J(i)} (\tilde{\varepsilon}_{k(j)} - \tilde{\varepsilon}_j)^2 / (2|J(i)|) - \tilde{\Sigma}_i| > Cn^{-c} \right).$$

By A3, $|f(X_{k(j)}) - f(X_j)| \leqslant L\|X_{k(j)} - X_j\| \leqslant 2Lb$. Since b converges to zero at a polynomial rate, $P_1 = 0$ if c and C in the definition of P_1 are sufficiently small and large, respectively. Consider P_3. Note that $P_3 \leqslant P_{31} + P_{32}$ where

$$P_{31} = \left(|\sum_{j \in J(i)} \tilde{\varepsilon}_j^2 / |J(i)| - \tilde{\Sigma}_i| > Cn^{-c} \right) \text{ and } P_{32} = \left(|\sum_{j \in J(i)} \tilde{\varepsilon}_{k(j)} \tilde{\varepsilon}_j| / |J(i)| > Cn^{-c} \right).$$

Since $|\Sigma_i - \Sigma_j| \leqslant C\|X_i - X_j\|$ and b is polynomially small, it follows that

$$P_{31} = P\left(|\sum_{j \in J(i)} (\tilde{\varepsilon}_j^2 - \tilde{\Sigma}_j)| / |J(i)| > Cn^{-c} \right).$$

Then Hoeffding inequality gives (see proposition 1.3.5 in [15])

$$P_{31} \leqslant 2 \exp\left(-Cn^{-c} |J(i)| / M^2 \right).$$

Therefore, $nP_{31} \leqslant Cn^{-c}$ if $\min_{i=1,...,n} |J(i)|/M^2 > cn^C$, which holds by assumption (iii).

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 $|\tilde{\varepsilon}_j| \leqslant M$ for all $j = 1, ..., n$, $nP_{32} \leqslant Cn^{-c}$ like $nP_{31} \leqslant Cn^{-c}$. Similar argument shows that $nP_2 \leqslant Cn^{-c}$ 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 $\|A^{1/2} - B^{1/2}\| \leqslant p^{1/2}\|A - B\|^{1/2}$.*

Comment 8. I was not able to find this result in the literature, so I provide a proof here.

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 $i = 1, ..., p$,

$$\sum_{j=1}^p (\lambda_i(A) - \lambda_j(B))^2 f_{ij}^2 = \|\sum_{j=1}^p (\lambda_i(A) - \lambda_j(B)) f_{ij} b_j\|^2$$

$$= \|\lambda_i(A) a_i - \sum_{j=1}^p \lambda_j(B) f_{ij} b_j\|^2 = \|(A - B)a_i\|^2 \leqslant \|A - B\|^2$$

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

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 $i = 1, ..., p$,

$$\|(A^{1/2} - B^{1/2})a_i\|^2 = \sum_{j=1}^p (\lambda_i^{1/2}(A) - \lambda_j^{1/2}(B))^2 f_{ij}^2 \leqslant \sum_{j=1}^p |\lambda_i(A) - \lambda_j(B)| f_{ij}^2$$

$$\leqslant \left(\sum_{j=1}^p (\lambda_i(A) - \lambda_j(B))^2 f_{ij}^2 \right)^{1/2} \leqslant \|A - B\|$$

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

$$\|(A^{1/2} - B^{1/2})c\| = \|(A^{1/2} - B^{1/2}) \sum_{i=1}^p d_i a_i\|$$

$$\leqslant \sum_{i=1}^p |d_i| \|(A^{1/2} - B^{1/2})a_i\| \leqslant \sum_{i=1}^p |d_i| \|A - B\|^{1/2} \leqslant p^{1/2}\|A - B\|^{1/2}$$

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

A.3. **Primitive Conditions for A1.** In this section, I give a counter-example for the statement that for A1 to hold, it suffices to assume that X_i's are sampled from a distribution that is absolutely continuous with respect to Lebesgue measure, has bounded support, and whose density is bounded from above and away from zero on the support. I also prove that A1 holds if, in addition to above conditions, one assumes that the support is a convex set.

Lemma 3. *There exists a probability distribution on $\mathbb{R}^2$ with bounded support such that this distribution is uniform on its support and if X_i's are sampled from this distribution, then A1 fails.*

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

$$S = \{(x_1, x_2) \in [0,1] \times [-(1+\alpha)/2, (1+\alpha)/2] : x_1 \geqslant 0;\ -(1+\alpha)x_1^\alpha/2 \leqslant x_2 \leqslant (1+\alpha)x_1^\alpha/2\}$$

for some $\alpha > 0$. For fixed i, the probability that $X_{i,1} \leqslant \underline{h}$ 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 $X_{i,1} \leqslant \underline{h}$ 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

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

Set $\underline{h} = (c/n)^{1/(1+\alpha)}$ and $\overline{h} = (C/n)^{1/(1+\alpha)}$ with $0 < c < C < 1$. Then I can find the limit of $P(A_n)$ as $n \to \infty$:

$$\lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} c(1 - C/n)^{n-1} = ce^{-C} > 0.$$

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 $(C^{1/(1+\alpha)} - c^{1/(1+\alpha)})/n^{1/(1+\alpha)}$. 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 A1.

Lemma 4. *Suppose that A4 holds. If X_i's are sampled from a distribution that is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$, 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 A1 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 $0 < C \leqslant 1$ 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 Lebesgue 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 $N(\delta) \leqslant C/\delta^d$. Consider the lower bound in A1(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 dropping into these covering balls. Since the density is bounded away from zero and from above, there exist some constants $c, C > 0$ such that for each $h \in H_n$ and $j = 1, ..., N(h)$, $ch^d \leqslant P(X_i \in B(G_{h,j}, \delta_h)) \leqslant Ch^d$. Denote $I_{h,j}(X_i) = I\{X_i \in B(G_{h,j}, \delta_h)\}$. Bernstein inequality (see proposition 1.3.2 in [15]) gives

$$P\left(\sum_{i=1}^n I_{h,j}(X_i)/n < ch^d/2\right) \leqslant P\left(\sum_{i=1}^n I_{h,j}(X_i)/n - E[I_{h,d}(X_i)] < -ch^d/2\right)$$
$$\leqslant C\exp(-cnh^d).$$

Then by the union bound and A4,

$$P\left(\cup_{h\in H_n, j=1,...,N(h)}\left\{\sum_{i=1}^n I_{h,j}(X_i)/n < C_1 h^d/2\right\}\right) \leqslant Ch_{\min}^{-d}\log n \exp(-cnh_{\min}^d),$$

where I have used inequality $|H_n| \leqslant C\log n$, which follows from $h_{\max}a^{|H_n|} \asymp (\log n/n)^{1/(3d)}$. By A4, $nh_{\min}^d > Cn^c$. So, summing the probabilities above over n, I conclude, by the Borel-Cantelli lemma, that the lower bound in A1(ii) holds for sufficiently large n a.s. A similar argument gives the upper bound. So, A1 holds. □

A.4. **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 [12]. 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. $\sup_{w\in\mathbb{R}} f_W(w) \leqslant C\sqrt{\log S}$ *for some universal constant* C.

Proof. Theorem 3 in [12] proves that $\sup_{w\in\mathbb{R}} f_W(w) \leqslant CE[W]$. In addition, it follows from the same argument as in Lemma 8 that $E[W] \leqslant C\sqrt{\log S}$. Combining these bounds gives the result. □

A.5. **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* $w_{i,n} \leqslant C\sqrt{\log n}$ *with* $C \in (0, 1)$ *for all* $i = 1, ..., n$, $n = 1, ..., \infty$, *then*

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

Proof. The proof is closely related to that in Lemma 6.2 in [16]. 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,

$$t_n^2 = \sum_{i=1}^n \left(E[Z_{i,n}^2] - (E[Z_{i,n}])^2\right)/n^2 \leqslant \sum_{i=1}^n \exp(w_{i,n}^2)/n^2 \to 0$$

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

$$E\left[\left|n^{-1}\sum_{i=1}^n \exp(w_{i,n}\xi_i - w_{i,n}^2/2) - 1\right|\right] = E\left[\left|\sum_{i=1}^n Z_{i,n}/n - 1\right|\right] \leqslant t_n \to 0.$$

The result follows. □

A.6. **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 conditions A1-A7 hold. I will use the following additional notation. Let $\{\psi_n\}_{n=1}^\infty$ be a sequence of positive real numbers such that $\psi_n \geqslant C_\psi (p \log n)^{1/2}/n^{c_\psi}$ for some sufficiently large $C_\psi > 0$ and sufficiently small $c_\psi > 0$ and $\psi_n \leqslant Cn^{-c}$ for all n. 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 $\widehat{\Sigma}_i$ 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 $\widehat{e}_j = \widehat{\Sigma}^{1/2}\epsilon_j$ and $e_j = \Sigma^{1/2}\epsilon_j$. Note that $\widehat{e}_j$ is equal in distribution to $\tilde{Y}_j \sim N(0_p, \widehat{\Sigma}_i)$. Finally, denote

$$\varepsilon_{(i,m,h)} = \sum_{j=1}^n w_h(X_i, X_j)\varepsilon_{j,m} \quad \text{and} \quad f_{(i,m,h)} = \sum_{j=1}^n w_h(X_i, X_j) f_m(X_j),$$

$$e_{(i,m,h)} = \sum_{j=1}^n w_h(X_i, X_j) e_{j,m} \quad \text{and} \quad \widehat{e}_{(i,m,h)} = \sum_{j=1}^n w_h(X_i, X_j)\widehat{e}_{j,m},$$

$$T^{PIA} = \max_{s \in S_n}(\widehat{e}_s/\widehat{V}_s) \quad \text{and} \quad T^{PIA,0} = \max_{s \in S_n}(e_s/V_s).$$

Note that T^{PIA} is equal in distribution to the simulated statistic for the plug-in critical value.

I start with a result on bounds for weights and variances of the kernel estimator. The same result can be found in [22]. I provide the proof here for convenience of the reader.

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

$$w_h(X_i, X_j) \leqslant C/(nh^d)$$

and

$$c/\sqrt{nh^d} \leqslant V_{(i,m,h)} \leqslant C/\sqrt{nh^d}$$

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

Proof. By A1 and A5, for any $i = 1, ..., n$ and $h \in H_n$,

$$cnh^d \leqslant cM_{h/2}(X_i) \leqslant \sum_{k=1}^{n} K(X_i - X_k) \leqslant CM_h(X_i) \leqslant Cnh^d$$

and

$$cnh^d \leqslant \sum_{k=1}^{n} K^2(X_i - X_k) \leqslant Cnh^d.$$

In addition, $K(X_i - X_j) \leqslant 1$ for any $j = 1, ..., n$, and so

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

By A2, since $\sum_{j=1}^{n} w_h(X_i, X_j) = 1$,

$$V_{(i,m,h)} = \left(\sum_{j=1}^{n} w_h^2(X_i, X_j) \Sigma_{j,mm} \right)^{1/2}$$

$$\leqslant C \left(\sum_{j=1}^{n} w_h^2(X_i, X_j) \right)^{1/2} \leqslant C \max_{j=1,...,n} w_h^{1/2}(X_i, X_j) \leqslant C/\sqrt{nh^d}$$

and

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

The claim of the lemma follows. □

Lemma 8. *$E[\max_{s \in S_n} |e_s/V_s|] \leqslant C(\log n)^{1/2}$ uniformly over the set of models $\mathcal{G}$. In particular, $c_{1-\lambda}^{PIA,0} \leqslant C\sqrt{\log n}/\lambda$ for all $\lambda \in (0, 1)$ uniformly over the set of models $\mathcal{G}$. In addition, $P(\max_{s \in S_n} |e_s/V_s| > C\sqrt{\log n}) \leqslant Cn^{-c}$ for sufficiently small and large constants c and C, respectively, 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 [37],

$$E \left[\max_{s \in S_n} |e_s/V_s| \right] \leqslant C \left\| \max_{s \in S_n} |e_s/V_s| \right\|_\psi \leqslant C(\log n)^{1/2}$$

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

$$\lambda \leqslant P \left(\max_{s \in S_n} |e_s/V_s| \geqslant c_{1-\lambda}^{PIA,0} \right) \leqslant E \left[\max_{s \in S_n} |e_s/V_s| \right] / c_{1-\lambda}^{PIA,0} \leqslant C\sqrt{\log n}/c_{1-\lambda}^{PIA,0}$$

for any $\lambda \in (0, 1)$. So, $c_{1-\lambda}^{PIA,0} \leqslant C\sqrt{\log n}/\lambda$. The third result follows from Borell inequality (see, for example, Proposition A.2.1 in [37]). □

Lemma 9. *$P(\max_{s\in S_n}|\widehat{V}_s/V_s-1|>Cn^{-c})\leqslant Cn^{-c}$ and $P(\max_{s\in S_n}|V_s/\widehat{V}_s-1|>Cn^{-c})\leqslant Cn^{-c}$ uniformly over the set of models $\mathcal{G}$.*

Proof. By A2, for any $(i,m,h)\in S_n$,

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

In addition,

$$\left|\widehat{V}^2_{(i,m,h)}-V^2_{(i,m,h)}\right|\leqslant\sum_{j=1}^n w_h^2(X_i,X_j)|\widehat{\Sigma}_{j,mm}-\Sigma_{j,mm}|.$$

So,

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

So,

$$P\left(\max_{s\in S_n}|\widehat{V}_s^2/V_s^2-1|>Cn^{-c}\right)\leqslant Cn^{-c}$$

by A6. Combining this result with inequality $|x-1|\leqslant|x^2-1|$, 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. *$P(c^{PIA,0}_{1-\lambda-\psi_n}>c^{PIA}_{1-\lambda})\leqslant Cn^{-c}$ and $P(c^{PIA,0}_{1-\lambda+\psi_n}<c^{PIA}_{1-\lambda})\leqslant Cn^{-c}$ uniformly over all $\lambda\in(0,1)$[16] and over the set of models $\mathcal{G}$ where ψ_n is defined in the beginning of this section ($\psi_n\geqslant C_\psi(p\log n)^{1/2}/n^{c_\psi}$ and $\psi_n\leqslant Cn^{-c}$).*

Proof. Denote

$$p_1=\max_{s\in S_n}\left|\frac{e_s}{V_s}\right|\max_{s\in S_n}\left|\frac{V_s}{\widehat{V}_s}-1\right|$$

and

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

where $(\cdot)_m$ denotes mth component of the vector $(\cdot)$. Then, by the triangle inequality,

$$|T^{PIA}-T^{PIA,0}|\leqslant p_1+p_2$$

by the triangle inequality. Let A denote the event $\{\max_{j=1,\ldots,n}\|\widehat{\Sigma}_j-\Sigma_j\|<C_4n^{-c_4}\}$. By A6, $P(A)\geqslant 1-Cn^{-c}$ as $n\to\infty$. Thus, it is enough to show that $c^{PIA,0}_{1-\lambda-\psi_n}\leqslant c^{PIA}_{1-\lambda}$ and $c^{PIA,0}_{1-\lambda+\psi_n}\geqslant c^{PIA}_{1-\lambda}$ on A.

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

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

$$P(p_1 > C\sqrt{\log n}n^{-c}B|Y_1^n) \leqslant 1/B$$

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 $m = 1, ..., p$,

$$\begin{aligned} &E\left[((\widehat{\Sigma}_j^{1/2} - \Sigma_j^{1/2})\epsilon_j)_m^2|Y_1^n\right] \leqslant E\left[\|(\widehat{\Sigma}_j^{1/2} - \Sigma_j^{1/2})\epsilon_j\|^2|Y_1^n\right] \\ &\leqslant E\left[\|\widehat{\Sigma}_j^{1/2} - \Sigma_j^{1/2}\|^2\|\epsilon_j\|^2|Y_1^n\right] \leqslant p\|(\widehat{\Sigma}_j^{1/2} - \Sigma_j^{1/2})\|^2 \leqslant p^2\|\widehat{\Sigma}_j - \Sigma_j\| \end{aligned}$$

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

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

is a mean-zero Gaussian random variable with variance bounded by Cp^2n^{-c} for any $(i, m, h) \in S_n$. In addition, on A, $\max_{s\in S_n} V_s/\widehat{V}_s \leqslant 2$ for sufficiently large n. Thus, Markov inequality and the argument like that used in Lemma 8 yield

$$P(p_2 > C\sqrt{\log n}pn^{-c}B|Y_1^n) \leqslant 1/B$$

on A. Let $B = Cn^c/(p\log n)^{1/2}$. Recall that $\psi_n \geqslant C_\psi(p\log n)^{1/2}/n^{c_\psi}$. Since c_ψ and C_ψ are assumed to be sufficiently small and large correspondingly, I can and will assume that $\psi_n \geqslant 4/B$. I will also assume that $\psi_n \geqslant C(p\log n)n^{-c}B$ (recall that c and C can change at each appearance).

Note that $T^{PIA,0}$ is the maximum over $|S_n|$ standard Gaussian random variables. Since $|S_n| \leqslant Cn^\phi$ for some $\phi > 0$, Lemma 5 gives $c^{PIA,0}_{1-\lambda-\psi_n/2} - c^{PIA,0}_{1-\lambda-\psi_n} \geqslant c\psi_n/(\log n)^{1/2}$, so that

$$c^{PIA,0}_{1-\lambda-\psi_n/2} - c^{PIA,0}_{1-\lambda-\psi_n} \geqslant C\sqrt{\log n}pn^{-c}B.$$

Now the first part of the lemma follows from

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

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

Lemma 11. *$|P(\max_{s\in S_n}(\varepsilon_s/V_s) \leqslant c^{PIA,0}_{1-\lambda}) - (1-\lambda)| \leqslant Cn^{-c}$ and $|P(-\max_{s\in S_n}(\varepsilon_s/V_s) \leqslant c^{PIA,0}_{1-\lambda}) - (1-\lambda)| \leqslant Cn^{-c}$ uniformly over all $\lambda \in (0,1)$ and over the set of models $\mathcal{G}$.*

Proof. By Lemma 7 and A2, for any $(i,m,h) \in S_n$ and $j = 1, ..., n$,

$$\Sigma_{i,mm}^{1/2} w_h(X_i, X_j)/V_{(i,m,h)} \leqslant C/\sqrt{nh^d} \leqslant C/\sqrt{nh_{\min}^d}.$$

Therefore, both claims of the lemma follows by combining A4 and Corollary 2.1, case E.4 in [13]. □

Lemma 12. *$P(\max_{s\in S_n} |\varepsilon_s/V_s| > C\sqrt{\log n}) \leqslant Cn^{-c}$ and $P(\max_{s\in S_n} |\varepsilon_s/\widehat{V}_s| > C\sqrt{\log n}) \leqslant Cn^{-c}$ for sufficiently small and large constants c and C, respectively, uniformly over the set of models $\mathcal{G}$.*

Proof. Set $\lambda = Cn^{-c}$ for sufficiently small and large c and C, respectively. Then Lemma 11 implies that

$$P\left(\max_{s\in S_n}(\varepsilon_s/V_s) > c_{1-Cn^{-c}}^{PIA,0}\right) \leqslant Cn^{-c}. \tag{21}$$

Further, Borell's inequality (see, for example, Proposition A.2.1 in [37]) implies that

$$c_{1-Cn^{-c}}^{PIA,0} \leqslant C\sqrt{\log n} \tag{22}$$

if C is sufficiently large. Combining (21) and (22) implies the first asserted claim. The second claim follows by noting that

$$\max_{s\in S_n} |\varepsilon_s/\widehat{V}_s| \leqslant \max_{s\in S_n} |\varepsilon_s/V_s| \max_{s\in S_n}(V_s/\widehat{V}_s)$$

and that $P(\max_{s\in S_n} |V_s/\widehat{V}_s| \leqslant 1 + Cn^{-c}) \geqslant 1 - Cn^{-c}$ by Lemma 9. □

Lemma 13. *$P(\max_{s\in S_n\setminus S_n^D} \widehat{f}_s/\widehat{V}_s > 0) \leqslant Cn^{-c}$ uniformly over the set of models $\mathcal{G}$.*

Proof. By Lemma 11,

$$\left|P\left(\max_{s\in S_n}(\varepsilon_s/V_s) \leqslant c_{1-\gamma_n-\psi_n}^{PIA,0}\right) - (1-\gamma_n-\psi_n)\right| \leqslant Cn^{-c}.$$

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

$$\begin{aligned}
P\left(\max_{s\in S_n\setminus S_n^D}(\widehat{f}_s/\widehat{V}_s) > 0\right) &= P\left(\max_{s\in S_n\setminus S_n^D}(\widehat{f}_s/V_s) > 0\right) \\
&= P\left(\max_{s\in S_n\setminus S_n^D}(f_s/V_s + \varepsilon_s/V_s) > 0\right) \\
&\leqslant P\left(\max_{s\in S_n\setminus S_n^D}(-c_{1-\gamma_n-\psi_n}^{PIA,0} + \varepsilon_s/V_s) > 0\right) \\
&\leqslant P\left(\max_{s\in S_n}(\varepsilon_s/V_s) > c_{1-\gamma_n-\psi_n}^{PIA,0}\right) \\
&\leqslant 1 - (1-\gamma_n-\psi_n) + Cn^{-c} \\
&= \gamma_n + \psi_n + Cn^{-c}.
\end{aligned}$$

Noting that $\gamma_n + \psi_n \leqslant Cn^{-c}$, which holds by the definition of ψ_n and A7, yields the result. □

Lemma 14. *$P(S_n^D \subset S_n^{RMS}) \geqslant 1 - Cn^{-c}$ uniformly over the set of models $\mathcal{G}$.*

Proof. By Lemma 10, $P(c_{1-\gamma_n-\psi_n}^{PIA,0} > c_{1-\gamma_n}^{PIA}) \leqslant Cn^{-c}$. In addition, for any $x \in (-1,1)$,

$$2/(1+x) - 1 \geqslant 2(1-x) - 1 \geqslant 1 - 2x \geqslant 1 - 2|x|.$$

So,

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

By Lemma 8, that $c_{1-\gamma_n-\psi_n}^{PIA,0} \leqslant C(\log n)^{1/2}/(\gamma_n+\psi_n)$. By Lemma 9, $P(|\max_{s\in S_n^D}(V_s/\widehat{V}_s) - 1| \leqslant Cn^{-c}) \geqslant 1 - Cn^{-c}$. So, with probability at least $1 - Cn^{-c}$,

$$c_{1-\gamma_n-\psi_n}^{PIA,0}\left(1 - 2|\max_{s\in S_n^D}(V_s/\widehat{V}_s) - 1|\right) \geqslant c_{1-\gamma_n-\psi_n}^{PIA,0} - C(\log n)^{1/2}n^{-c}/(\gamma_n+\psi_n).$$

Take $\chi_n = C(\log n)n^{-c}/(\gamma_n+\psi_n)$. Then $\chi_n \leqslant Cn^{-c}$ by the choice of ψ_n (recall that the constant c_ψ in the definition of ψ_n is sufficiently small). By Lemma 5,

$$c_{1-\gamma_n-\psi_n}^{PIA,0} - C(\log n)^{1/2}n^{-c}/(\gamma_n+\psi_n) \geqslant c_{1-\gamma_n-\psi_n-\chi_n}^{PIA,0}. \tag{23}$$

Indeed, by Lemma 5, the density of $\max_{s\in S_n}(e_s/V_s)$ is bounded from above by $C(\log n)^{1/2}$ where $C > 0$ is a universal constant, so that $C(\log n)^{1/2}(c_{1-\gamma_n-\psi_n}^{PIA,0} - c_{1-\gamma_n-\psi_n-\chi_n}^{PIA,0}) \geqslant \chi_n$; hence (23) follows from the definition of χ_n. Therefore,

$$\begin{aligned}
P\left(S_n^D \subset S_n^{RMS}\right) &\geqslant P\left(\max_{s\in S_n}(-\varepsilon_s/V_s) < c_{1-\gamma_n-\psi_n-\chi_n}^{PIA,0}\right) - Cn^{-c} \\
&\geqslant 1 - \gamma_n - \psi_n - \chi_n - Cn^{-c}.
\end{aligned}$$

The result follows since $\gamma_n + \psi_n + \chi_n \leqslant Cn^{-c}$ by the definitions of ψ_n and χ_n and A7. □

Lemma 15. *$P(S_n^{RMS} = S_n) \geqslant 1 - Cn^{-c}$ uniformly over the set of models $\mathcal{G}_{00}$.*

Proof. By Lemma 10, $P(c^{PIA,0}_{1-\gamma_n-\psi_n} > c^{PIA}_{1-\gamma_n}) \leqslant Cn^{-c}$. By Lemma 9, $P(\max_{s\in S_n}(V_s/\widehat{V}_s) \leqslant 1 + Cn^{-c}) \geqslant 1 - Cn^{-c}$. If $f = 0_p$, then for any $s \in S_n$, $\widehat{f}_s = \varepsilon_s$. So,

$$\begin{aligned}
P\left(S^{RMS}_n = S_n\right) &= P\left(\min_{s\in S_n}(\varepsilon_s/\widehat{V}_s) > -2c^{PIA}_{1-\gamma_n}\right) \\
&\geqslant P\left(\min_{s\in S_n}(\varepsilon_s/\widehat{V}_s) > -2c^{PIA,0}_{1-\gamma_n-\psi_n}\right) - Cn^{-c} \\
&\geqslant P\left(\min_{s\in S_n}(\varepsilon_s/V_s)\max_{s\in S_n}(V_s/\widehat{V}_s) > -2c^{PIA,0}_{1-\gamma_n-\psi_n}\right) - Cn^{-c} \\
&\geqslant P\left(\min_{s\in S_n}(\varepsilon_s/V_s)(1+Cn^{-c}) > -2c^{PIA,0}_{1-\gamma_n-\psi_n}\right) - Cn^{-c} \\
&\geqslant P\left(\min_{s\in S_n}(\varepsilon_s/V_s) > -2c^{PIA,0}_{1-\gamma_n-\psi_n}(1-Cn^{-c})\right) - Cn^{-c} \\
&\geqslant P\left(\min_{s\in S_n}(\varepsilon_s/V_s) > -c^{PIA,0}_{1-\gamma_n-\psi_n}\right) - Cn^{-c} \\
&= P\left(\max_{s\in S_n}(-\varepsilon_s/V_s) < c^{PIA,0}_{1-\gamma_n-\psi_n}\right) - Cn^{-c}
\end{aligned}$$

where I have used the inequality $c^{PIA,0}_{1-\gamma_n-\psi_n} \geqslant 0$ in the third line and where the line preceding the last one follows from $2(1-Cn^{-c}) \geqslant 1$ for large n. Combining these results with Lemma 11 yields

$$P\left(S^{RMS}_n = S_n\right) \geqslant 1 - \gamma_n - \psi_n - Cn^{-c}.$$

The result follows by noting that $\gamma_n + \psi_n \leqslant Cn^{-c}$. □

Lemma 16. *$P(c^{PIA}_{1-\alpha} > C\sqrt{\log n}) \leqslant Cn^{-c}$ and $P(c^{RMS}_{1-\alpha} > C\sqrt{\log n}) \leqslant Cn^{-c}$ for sufficiently small and large c and C, respectively, uniformly over the set of models $\mathcal{G}$.*

Proof. Since $S^{RMS}_n \subseteq S_n$, it follows that $c^{RMS}_{1-\alpha} \leqslant c^{PIA}_{1-\alpha}$. Therefore, the second claim follows from the first one. To prove the first claim, note that $c^{PIA,0}_{1-\alpha+\psi_n} \leqslant c^{PIA,0}_{1-\alpha/2}$ for large n, and so Lemma 10 implies that $P(c^{PIA,0}_{1-\alpha/2} < c^{PIA}_{1-\alpha}) \leqslant Cn^{-c}$. In addition $c^{PIA,0}_{1-\alpha/2} \leqslant C\sqrt{\log n}$ by Lemma 8. Combining these results yields the asserted claim. □

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 $g \in \mathcal{F}(\tau, L)$. Then $\partial g(x_1, ..., x_d)/\partial x_m \geqslant 0$ for all $m = 1, ..., d$ implies that for any $y = (y_1, ..., y_d) \in \mathbb{R}^d$ satisfying $0 \leqslant y \leqslant h$,*

$$g(x+y) - g(x) \geqslant -\frac{\max(L^{\tau-[\tau]}, L)}{\prod_{j=1,...,[\tau]}(\tau - [\tau] + j)}\|h\|^\tau.$$

Proof. For any $y = (y_1, ..., y_d) \in \mathbb{R}^d$ satisfying $0 \leqslant y \leqslant h$, let $l = y/\|y\|$. Then $g^{(1,l)}(x) \geqslant 0$. If $g^{(1,l)}(x + tl) \geqslant 0$ 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 $k = 1, ..., [\tau]$. Then $g^{([\tau],l)}(x + tl) \geqslant -(L(t - t_0))^{\tau - [\tau]}$. Integrating it $[\tau]$ times gives

$$g(x + y) - g(x) \geqslant -\frac{L^{\tau-[\tau]}}{\prod_{j=1,...,[\tau]}(\tau - [\tau] + j)}\|y\|^{\tau}. \tag{24}$$

The asserted claim follows by noting that $\|y\| \leqslant \|h\|$. □

A.7. **Proofs of Theorems.**

Proof of Theorem 1. Consider any $w \in \mathcal{G}_0$. For any $s \in S_n$, $f_s \leqslant 0$ since the kernel K is positive by A5. By Lemma 10, $P(c^{PIA,0}_{1-\alpha-\psi_n} > c^{PIA}_{1-\alpha}) \leqslant Cn^{-c}$. By Lemma 9, $P(\max_{s\in S_n}(V_s/\widehat{V}_s) \leqslant 1 + Cn^{-c}) \geqslant 1 - Cn^{-c}$. So,

$$\begin{aligned}
P\left(\widehat{T} \leqslant c^{PIA}_{1-\alpha}\right) &= P\left(\max_{s\in S_n}(\widehat{f}_s/\widehat{V}_s) \leqslant c^{PIA}_{1-\alpha}\right) \\
&\geqslant P\left(\max_{s\in S_n}(\varepsilon_s/\widehat{V}_s) \leqslant c^{PIA}_{1-\alpha}\right) \\
&\geqslant P\left(\max_{s\in S_n}(\varepsilon_s/\widehat{V}_s) \leqslant c^{PIA,0}_{1-\alpha-\psi_n}\right) - Cn^{-c} \\
&\geqslant P\left(\max_{s\in S_n}(\varepsilon_s/V_s)\max_{s\in S_n}(V_s/\widehat{V}_s) \leqslant c^{PIA,0}_{1-\alpha-\psi_n}\right) - Cn^{-c} \\
&\geqslant P\left(\max_{s\in S_n}(\varepsilon_s/V_s)(1 + Cn^{-c}) \leqslant c^{PIA,0}_{1-\alpha-\psi_n}\right) - Cn^{-c}
\end{aligned}$$

where I have used the inequality $c^{PIA,0}_{1-\alpha-\psi_n} \geqslant 0$ in the fourth line. Let $\chi_n = C(\log n)n^{-c}$. Since $P(\max_{s\in S_n}|\varepsilon_s/V_s| > C\sqrt{\log n}) \leqslant Cn^{-c}$ for sufficiently small and large c and C, respectively, by Lemma 12, an application of Lemma 5 shows that the last expression is bounded from below by

$$P\left(\max_{s\in S_n}(\varepsilon_s/V_s) \leqslant c^{PIA,0}_{1-\alpha-\psi_n-\chi_n}\right) - Cn^{-c}.$$

Here the application of Lemma 5 is the same as in the proof of Lemma 14. Then $P(\widehat{T} \leqslant c^{PIA}_{1-\alpha}) \geqslant 1 - \alpha - Cn^{-c}$ follows from this bound and Lemma 11 since $\psi_n + \chi_n \leqslant Cn^{-c}$.

Now consider the RMS critical value. By Lemma 14, $P(c^{D}_{1-\alpha} > c^{RMS}_{1-\alpha}) \leqslant Cn^{-c}$. By Lemma 13, $P(\max_{s\in S_n\setminus S^D_n}\widehat{f}_s/\widehat{V}_s > 0) \leqslant Cn^{-c}$. So,

$$\begin{aligned}
P\left(\widehat{T} \leqslant c^{RMS}_{1-\alpha}\right) &= P\left(\max_{s\in S_n}(\widehat{f}_s/\widehat{V}_s) \leqslant c^{RMS}_{1-\alpha}\right) \\
&\geqslant P\left(\max_{s\in S_n}(\widehat{f}_s/\widehat{V}_s) \leqslant c^{D}_{1-\alpha}\right) - Cn^{-c} \\
&\geqslant P\left(\max_{s\in S^D_n}(\widehat{f}_s/\widehat{V}_s) \leqslant c^{D}_{1-\alpha}\right) - Cn^{-c}.
\end{aligned}$$

Since S_n^D is non-stochastic, 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 $w \in \mathcal{G}_{00}$ so that $f = 0_p$. By Lemma 10, $P(c_{1-\alpha+\psi_n}^{PIA,0} < c_{1-\alpha}^{PIA}) \leqslant Cn^{-c}$. By Lemma 9, $P(\min_{s\in S_n}(V_s/\widehat{V}_s) \geqslant 1 - Cn^{-c}) \geqslant 1 - Cn^{-c}$. So,

$$\begin{aligned}
P\left(\widehat{T} \leqslant c_{1-\alpha}^{PIA}\right) &= P\left(\max_{s\in S_n}(\widehat{f}_s/\widehat{V}_s) \leqslant c_{1-\alpha}^{PIA}\right) \\
&= P\left(\max_{s\in S_n}(\varepsilon_s/\widehat{V}_s) \leqslant c_{1-\alpha}^{PIA}\right) \\
&\leqslant P\left(\max_{s\in S_n}(\varepsilon_s/\widehat{V}_s) \leqslant c_{1-\alpha+\psi_n}^{PIA,0}\right) + Cn^{-c} \\
&\leqslant P\left(\max_{s\in S_n}(\varepsilon_s/V_s)\min_{s\in S_n}(V_s/\widehat{V}_s) \leqslant c_{1-\alpha+\psi_n}^{PIA,0}\right) + Cn^{-c} \\
&\leqslant P\left(\max_{s\in S_n}(\varepsilon_s/V_s)(1 - Cn^{-c}) \leqslant c_{1-\alpha+\psi_n}^{PIA,0}\right) + Cn^{-c}.
\end{aligned}$$

An argument like that used above shows that the last expression is bounded from above by $1 - \alpha + Cn^{-c}$.

For the RMS critical value, note that by Lemma 15, $P(S_n^{RMS} = S_n) \geqslant 1 - Cn^{-c}$ whenever $f = 0_p$. So,

$$P\left(\widehat{T} \leqslant c_{1-\alpha}^{RMS}\right) = P\left(\widehat{T} \leqslant c_{1-\alpha}^{PIA}\right) + Cn^{-c} \leqslant 1 - \alpha + Cn^{-c}.$$

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

Proof of Theorem 2. For any $w \in \mathcal{G}_\rho$, there exist $i \in \mathbb{N}$ and $m = 1, ..., p$ such that $f_m(X_i) \geqslant 3\rho/4$. By A3, there exists a ball $B_\delta(X_i)$ with center at X_i and radius δ such that $f_m(X_j) \geqslant \rho/2$ 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 $n \geqslant N$, there exists a triple $s_n = (i_n, m, h_n) \in S_n$ with h_n bounded away from zero such that $f_m(X_j) \geqslant \rho/2$ for all $X_j \in B_{h_n}(X_{i_n})$. Hence, $f_{s_n} \geqslant \rho/2$. Lemma 7 gives $V_{s_n} \leqslant Cn^{-\phi}$ for some $\phi > 0$, so $f_{s_n}/V_{s_n} > cn^\phi$. By Lemma 9, $P(|\widehat{V}_{s_n}/V_{s_n} - 1| > Cn^{-c}) \leqslant Cn^{-c}$. So, $P\{f_{s_n}/\widehat{V}_{s_n} > cn^\phi\} \geqslant 1 - Cn^{-c}$ for sufficiently small $c > 0$. Thus,

$$\begin{aligned}
P\left(\widehat{T} \leqslant c_{1-\alpha}^{P}\right) &\leqslant P\left(f_{s_n}/\widehat{V}_{s_n} \leqslant c_{1-\alpha}^{P} + \max_{s\in S_n}|\varepsilon_s/\widehat{V}_s|\right) \\
&\leqslant P\left(c_{1-\alpha}^{P} + \max_{s\in S_n}|\varepsilon_s/\widehat{V}_s| > cn^\phi\right) + Cn^{-c}
\end{aligned}$$

where in the first line I have used the following chain of inequalities: $\widehat{T} \geqslant f_{s_n}/\widehat{V}_{s_n} + \varepsilon_{s_n}/\widehat{V}_{s_n} \geqslant f_{s_n}/\widehat{V}_{s_n} - \max_{s\in S_n}|\varepsilon_{s_n}/\widehat{V}_{s_n}|$. The result follows by noting that from Lemmas 12 and 16, $P(c_{1-\alpha}^{P} + \max_{s\in S_n}|\varepsilon_s/\widehat{V}_s| > C\sqrt{\log n}) \leqslant Cn^{-c}$. □

Proof of Theorem 3. Let $f^0 = f(w_0)$ and for all $n \geqslant 1$, $f^n = f(w_n)$. As in the proof of Theorem 2, since $\rho(w_0, H_0) > 0$, there exists $i \in \mathbb{N}$ such that $f^0_m(X_i) \geqslant 3\rho/4$ for some $m = 1, ..., p$ and $\rho > 0$. In addition, by A3, there exists a ball $B_\delta(X_i)$ such that $f^0_m(X_j) \geqslant \rho/2$ for all $X_j \in B_\delta(X_i)$. So, for some $N \in \mathbb{N}$ and any $n \geqslant N$, there exists a triple $s_n = (i_n, m, h_n) \in S_n$ with h_n bounded away from zero such that $f^0_m(X_j) \geqslant \rho/2$ for all $X_j \in B_{h_n}(X_{i_n})$. Hence, $f^n_{s_n} \geqslant a_n\rho/2$. By Lemma 7, $V_{s_n} \leqslant C/\sqrt{n}$. Then Lemma 9 gives $P(f^n_{s_n}/\widehat{V}_{s_n} > ca_n/\sqrt{n}) \to 1$. The same argument as in the proof of Theorem 2 yields

$$P\left(\widehat{T} \leqslant c^P_{1-\alpha}\right) \leqslant P\left(c^P_{1-\alpha} + \max_{s\in S_n} |\varepsilon_s/\widehat{V}_s| > ca_n\sqrt{n}\right) + o(1).$$

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

Proof of Theorem 4. First, consider $\tau \leqslant 1$ case. Since $d \geqslant 1$, I have $\tau \leqslant d$. Consider any $w \in \mathcal{G}_\vartheta$. Since $\inf_{w\in\mathcal{G}_\vartheta} \rho_\vartheta(w, H_0)/h^\tau_{\min} \to \infty$, there exists a sequence a_n of positive numbers such that $a_n \to \infty$ and $\rho_\vartheta(w, H_0) > a_n h^\tau_{\min}$, and so there exist $i \in \mathbb{N}_\vartheta$ and $m = 1, ..., p$ such that $f_m(X_i) \geqslant a_n h^\tau_{\min}$. Let $s_n(w) = (i, m, h_{\min}) \in S_n$. By A3, $f_m(X_l) \geqslant ca_n h^\tau_{\min}$ for all $l = 1, ..., n$ such that $X_l \in B_{h_{\min}}(X_i)$. So, $f_{s_n(w)} \geqslant ca_n h^\tau_{\min}$. By A4, $nh^{3d}_{\min}/\log n \geqslant c$. By Lemma 7, $V_{s_n(w)} \leqslant C/\sqrt{nh^d_{\min}}$. So,

$$f_{s_n(w)}/\left(V_{s_n(w)}\sqrt{\log n}\right) \geqslant ca_n\sqrt{nh^{2\tau+d}_{\min}/\log n} \geqslant ca_n\sqrt{nh^{3d}_{\min}/\log n} \to \infty$$

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 $\tau \leqslant d$. For any $w \in \mathcal{G}_\vartheta$, there exist $i \in \mathbb{N}_\vartheta$ and $m = 1, ..., p$ such that $f_m(X_i) \geqslant a_n h^\tau_{\min}$ where a_n is as defined above. For $m = 1, ..., d$, set $e_m = 2h_{\min}$ if $\partial f_m(X_i)/\partial x_m \geqslant 0$ and $-2h_{\min}$ 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 $x \in \mathcal{C}$, $f_m(x) \geqslant ca_n h^\tau_{\min}$. By the definition of $\mathbb{N}_\vartheta$ and A1, there exists $l = 1, ..., n$ such that $X_l \in B_{h_{\min}}(X_{i,1} + e_1, ..., X_{i,d} + e_d)$. Let $s_n(w) = (l, m, h_{\min}) \in S_n$. Then $f_{s_n(w)} \geqslant ca_n h^\tau_{\min}$. The rest of the proof follows from the same argument as in the case $\tau \leqslant 1$.

Suppose $\tau > d$. 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 $C(\log n/n)^{1/(2\tau+d)}$. For any $w \in \mathcal{G}_\vartheta$, the same construction as above gives $s_n(w) = (l, m, h_o) \in S_n$ such that $f_m(X_j) \geqslant \rho_\vartheta(w, H_0) - Ch^\tau_o$ for all $j = 1, ..., n$ such that $X_j \in B_{h_o}(X_l)$. Since $\rho_\vartheta(w, H_0) \geqslant a_n(\log n/n)^{\tau/(2\tau+d)}$ for some sequence of real numbers a_n such that $a_n \to \infty$ as $n \to \infty$, $f_{s_n(w)} \geqslant (a_n - C)(\log n/n)^{\tau/(2\tau+d)}$. By Lemma 7, $V_{s_n(w)} \leqslant C/\sqrt{nh^d_o}$. Then

$$f_{s_n(w)}/\left(V_{s_n(w)}\sqrt{\log n}\right) \geqslant c(a_n - C) \to \infty.$$

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

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

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 $h \in \mathbb{R}_+$, $v(\cdot/h, h) \in \mathcal{F}(\tau, L)$ 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 $h \in \mathbb{R}_+$, $q(\cdot, h) \in \mathcal{F}(\tau, L)$, $q(x,h) = 0$ 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 $m = 1, ..., p$, $f_m^l \in \mathcal{F}(\tau, L)$. 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 as a model with the regression function f^l, disturbances $\{\varepsilon_i\}_{i=1}^\infty$ and design points $\{X_i\}_{i=1}^\infty$. Note that $\rho_\vartheta(w_l, H_0) \geqslant C r_n$ for all $l = 1, ..., N_n$ for some constant C. In addition, let w_0 be a model with zero regression function, disturbances $\{\varepsilon_i\}_{i=1}^\infty$ and design points $\{X_i\}_{i=1}^\infty$.

As in the proof of Lemma 6.2 in [16], for any sequence $\phi_n = \phi_n(Y_1, ..., Y_n)$ of tests with $\sup_{w \in \mathcal{G}_0 \cap \mathcal{G}_X} E_w[\phi_n] \leqslant \alpha + o(1)$,

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

where the first line follows from $\inf_{w\in\mathcal{G}_X,\rho_\vartheta(w,H_0)\geqslant Cr_n} E_w[\phi_n] \leqslant \min_{l=1,...,N_n} E_{w_l}[\phi_n]$ and $E_{w_0}[\phi_n] \leqslant \alpha + o(1)$. Further,

$$\sum_{i=1}^{N_n} E_{w_l}[\phi_n]/N_n - E_{w_0}[\phi_n] + o(1) \leqslant E_{w_0}\left[\left(\sum_{i=1}^{N_n}(dP_{w_l}/dP_{w_0})/N_n - 1\right)\phi_n\right] + o(1)$$
$$\leqslant E_{w_0}\left[\left|\sum_{i=1}^{N_n}(dP_{w_l}/dP_{w_0})/N_n - 1\right|\right] + o(1)$$

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

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

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

$$E_{w_0}\left[\left|\sum_{i=1}^{N_n}(dP_{w_l}/dP_{w_0})/N_n - 1\right|\right] \to 0$$

if $Cn^{1/2}h_n^{\tau+d/2} < \tilde{C}(\log N_n)^{1/2}$ 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 $\log N_n \geqslant C\log n$ for some constant C. □

Proof of Corollary 1. The proof follows from the same arguments, line by line, as those used in the proof of Theorem 1. Condition $p_n \log n \leqslant C_6 n^{c_6}$ for some sufficiently small and large c_6 and C_6 is required to make sure that one can define a sequence ψ_n such that $\psi_n \geqslant C_\psi(p_n\log n)^{1/2}/n^{c_\psi}$ for some sufficiently small and large c_ψ and C_ψ, respectively, and $\psi_n \leqslant Cn^{-c}$. □

Proof of Corollary 2. To prove the first result, note that $f_m(X_i, Z_i) \leqslant Ca_n$ for all $i \in N$ and $m = 1, ..., p$. So, $f_s \leqslant Ca_n$ for any $s \in S_n$. Therefore, combining Lemmas 7 and 9 gives

$$P\left(\max_{s\in S_n}(f_s/\widehat{V}_s) \leqslant Ca\sqrt{n_a h_{\max}^{d_1}}\right) \geqslant 1 - Cn^{-c}.$$

Since $a_n\sqrt{n_a h_{\max}^d \log n} \leqslant Cn^{-c}$, 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

$$P\left(\widehat{T} \leqslant c_{1-\alpha}^P\right) \geqslant P\left(\max_{s\in S_n}(\varepsilon_s/V_s) \leqslant c_{1-\alpha}^P\right) - Cn^{-c} = 1 - \alpha + Cn^{-c}$$

for $P = PIA$ or RMS.

The second result follows from the same argument as in the proof of Theorem 1 since $-Ca_n \leqslant f_s \leqslant Ca_n$ ensures that the bias is again asymptotically negligible in the comparison with the concentration rate of the test statistic.

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 $f_m(X_j) \geqslant$

$\rho/2$ for all $X_j \in B_{h_n}(X_{i_n})$ and $\|X_{i_n}^2 - x_0\| \leqslant a_n$. The rest of the proof follows from the argument similar to that used in the proof of Theorem 2. □

References

[1] Andrews, D. W. K., and Guggenberger, P. (2009). Validity of Subsampling and Plug-in Asymptotic Inference for Parameters Defined by Moment Inequalities. *Econometric Theory*, **25**, 669-709.

[2] Andrews, D. W. K., and Han, S. (2009). Invalidity of the Bootstrap and m out of n Bootstrap for Interval Endpoints. *Econometrics Journal*, **12**, 5172-5199.

[3] Andrews, D. W. K., and Shi, X. (2013). Inference Based on Conditional Moment Inequalities. *Econometrica*, **81**, 609-666.

[4] Andrews, D. W. K., and Shi, X. (2011). Nonparametric Inference Based on Conditional Moment Inequalities. *Journal of Econometrics, forthcoming.*

[5] Andrews, D. W. K., and Soares, G. (2010). Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection. *Econometrica*, **78**, 119-157.

[6] Armstrong, T. (2011). Asymptotically Exact Inference in Conditional Moment Inequalities Models. *Unpublished manuscript.*

[7] Armstrong, T. (2011). Weighted KS Statistics for Inference on Conditional Moment Inequalities. *Unpublished manuscript.*

[8] Armstrong, T. (2012). On the Asymptotic Distribution of Variance Weighted KS Statistics. *Unpublished manuscript.*

[9] Bugni, F. (2010). Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set. *Econometrica*, **78**, 735-753.

[10] Canay, I. (2010). EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity. *Journal of Econometrics*, **156**, 408-425.

[11] Chernozhukov, V., Hong, H., and Tamer, E. (2007). Estimation and Confidence Regions for Parameter Sets in Econometric Models. *Econometrica*, **75**, 1243-1284.

[12] Chernozhukov, V., Chetverikov, D., and Kato, K. (2011). Comparison and Anti-Concentration Bounds for Maxima of Gaussian Random Vectors. *arXiv:1301.4807.*

[13] Chernozhukov,V., Chetverikov, D., and Kato, K. (2012). Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors. *The Annals of Statistics, forthcoming.*

[14] Chernozhukov, V., Lee, S. and Rosen, A. (2013). Intersection Bounds: Estimation and Inference. *Econometrica*, **81**, 667-737.

[15] Dudley, R. (1999). Uniform Central Limit Theorems. *Cambridge Studies in Advanced Mathematics.*

[16] Dumbgen, L., and Spokoiny, V. (2001). Multiscale Testing of Qualitative Hypotheses. *The Annals of Statistics*, **29**, 124-152.

[17] Fan, J., and Yao, Q. (1998). Efficient Estimation of Conditional Variance Functions in Stochastic Regression. *Biometrika*, **85**, 645-660.

[18] Gine, E., and Nickl, R. (2010). Confidence Bands in Density Estimation. *The Annals of Statistics*, **38**, 1122-1170.

[19] Haile, P., and Tamer, E. (2003). Inference with an Incomplete Model of English Auctions. *Journal of Political Economy*, **111**, 1-51.

[20] Hall, P. (1991). On Convergence Rates of Suprema. *Probability Theory and Related Fields*, **89**, 447-455.

[21] Hardle, W., and Tsybakov, A. (2007). Local Polinomial Estimators of the Volatility Function in Nonparametric Autoregression. *Journal of Econometrics*, **81**, 233-242.

[22] Horowitz, J. L., and Spokoiny, V. (2001). An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model against a Nonparametric Alternative. *Econometrica*, **69**, 599-631.

[23] Khan, S., and Tamer, E. (2009). Inference on Endogenously Censored Regression Models Using Conditional Moment Inequalities. *Journal of Econometrics*, **152**, 104-119.

[24] Kim, K. (2008). Set Estimation and Inference with Models Characterized by Conditional Moment Inequalities. *Unpublished manuscript.*

[25] Kress, R. (1999). Linear Integral Equations. *Springer.*

[26] Lee, S., Song, K., and Whang, Y. (2011). Testing function inequalities. *CEMMAP working paper CWP 12/11.*

[27] Lepski, O. and Spokoiny, V. (1999). Minimax Nonparametric Hypothesis Testing: the Case of an Inhomogeneous Alternative. *Bernoulli*, **5**, 333-358.

[28] Manski, C. and Tamer, E. (2002). Inference on Regressions with Interval Data on a Regressor or Outcome. *Econometrica*, **70**, 519-546.

[29] Milgrom, P. and Weber, R. (1982). A Theory of Auctions and Competitive Bidding. *Econometrica*, **50**, 1089-1122.

[30] Muller, H. and Stadtmuller, U. (1987). Estimation of Heteroscedasticity in Regression Analysis. *The Annals of Statistics*, **15**, 610-625.

[31] Pakes, A. (2010). Alternative Models for Moment Inequalities. *Econometrica*, **78**, 1783-1822.

[32] Rice, J. (1984). Bandwidth Choice for Nonparametric Kernel Regression. *The Annals of Statistics*, **12**, 1215-1230.

[33] Romano, J. and Shaikh, A. (2010). Inference for the Identified Sets in Partially Identified Econometric Models. *Econometrica*, **78**, 169-211.

[34] Romano, J. and Shaikh, A. (2008). Inference for Identifiable Parameters in Partially Identified Econometric Models. *Journal of Statistical Planning and Inference*, **138**, 2786-2807.

[35] Rosen, A. (2008). Confidence Sets for Partially Identified Parameters That Satisfy a Finite Number of Moment Inequalities. *Journal of Econometrics*, **146**, 107-117.

[36] Tsybakov, A. (2009). Introduction to Nonparametric Estimation. *Springer*.

[37] Van der Vaart, A. and Wellner, J. (1996). Weak Convergence and Empirical Processes with Applications to Statistics. *Springer*.