Econometric Theory, 2017, Page 1 of 42.
doi:10.1017/S0266466617000184

ADAPTIVE TESTS OF CONDITIONAL MOMENT INEQUALITIES

DENIS CHETVERIKOV
UCLA

Many economic models yield conditional moment inequalities that can be used for inference on parameters of these models. In this paper, I construct new tests of parameter hypotheses in conditional moment inequality models based on studentized kernel estimates of moment functions. The tests automatically adapt to the unknown smoothness of the moment functions, have uniformly correct asymptotic size, and are 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 methods only allow 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 tests developed in this paper are consistent and those tests are not.

1. INTRODUCTION

Models defined by conditional moment inequalities play an important role in econometrics. For example, such models are often used to study behavioral choice; see Pakes (2010) for a survey. In addition, they appear in estimation problems with interval data and problems with censoring; see Manski and Tamer (2002). These models also offer a convenient way to study treatment effects in randomized experiments as described in Lee, Song, and Whang (2013). In this paper, I develop a new method of testing hypotheses about parameters in these models.[1] The method gives tests that are adaptive and rate-optimal against certain classes of smooth alternatives, which implies that the tests have sound power properties against these alternatives. The tests are straightforward to implement and computationally simple. By inverting the tests, the method can be used for constructing confidence bands for the parameters in these models. The tests are robust to weakly and/or partially identified models.

I consider a model in which the parameter $\theta \in \Theta$ satisfies the following conditional moment inequalities:

$$\mathrm{E}[m_j(X,W,\theta) \mid X] \leqslant 0, \quad \text{for all } j = 1,\dots,p, \tag{1}$$

where $m_j : \mathbb{R}^d \times \mathbb{R}^w \times \Theta \to \mathbb{R}$, $j = 1,\dots,p$, are some known functions, X and W are random vectors in $\mathbb{R}^d$ and $\mathbb{R}^w$, respectively, and the inequalities hold almost

I thank Victor Chernozhukov for his guidance, numerous discussions and permanent support. I am also grateful to Isaiah Andrews, Tim Armstrong, Jin Hahn, Jerry Hausman, Kengo Kato, Anton Kolotilin, Simon Lee, Rosa Matzkin, Anna Mikusheva, and Adam Rosen for useful comments and discussions. Address correspondence to Denis Chetverikov, Economics Department, UCLA, Los Angeles, CA, USA; e-mail: chetverikov@econ.ucla.edu.

surely.[2] In this model, I am interested in testing the null hypothesis, H_0, that $\theta = \theta_0$ against the alternative, H_a, that $\theta \neq \theta_0$ using a random sample $(X_i, W_i)_{i=1}^n$ from the distribution of (X, W).

I use a test statistic that is based on kernel estimates of conditional moment functions $x \mapsto \mathrm{E}[m_j(X, W, \theta_0) \mid X = x]$ with many different bandwidth values h. 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, that is, each estimate is divided by its estimated standard deviation. The test statistic is formed as the maximum of these studentized estimates over x and h, and large values of the statistic 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 $m(X_i, W_i, \theta_0) - \mathrm{E}[m(X_i, W_i, \theta_0) \mid X_i]$, $i = 1, \dots, n$, only via second moments of the noise. For reasons similar to those discussed in Chernozhukov, Hong, and Tamer (2007) and Andrews and Soares (2010), the distribution of the test statistic in large samples depends heavily on the extent to which the inequalities are binding. Moreover, the parameters that measure to what extent the inequalities 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 Andrews and Shi (2013), 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. It is substantially different from the proof techniques used in Andrews and Shi (2013) and Chernozhukov, Lee, and Rosen (2013b). One of the advantages of the proof technique used in this paper is that it gives an explicit bound on the bootstrap approximation error of the distribution of the test statistic. In particular, it allows me to show that for the tests developed in this paper, the probability of rejecting H_0 when H_0 is true can exceed nominal level only by a *polynomially* (in n) small term. 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, Hall (1991).

The tests in this paper are inspired by the results developed in Andrews and Shi (2013), who also constructed tests for the null hypothesis $\theta = \theta_0$ against the alternative $\theta \neq \theta_0$ in the model (1). The major difference between the tests in this paper and the tests in that paper is that I do not truncate the variance of the kernel estimators below from zero. This important modification eventually leads to adaptive tests with an optimal rate of uniform consistency against certain classes of smooth alternatives. Implementing this modification and demonstrating that new tests still control asymptotic size requires developing new proof techniques, as discussed above.

Using conditional moment inequalities (1) for inference is difficult because these inequalities typically do not identify the parameter θ. Let

$$\Theta_I = \left\{\theta_0 \in \Theta \colon \mathrm{E}[m_j(X, W, \theta_0) \mid X] \leqslant 0 \text{ almost surely for all } j = 1, \ldots, p\right\}$$

denote the identified set of the values θ_0 of the parameter θ that are consistent with the model (1). It is said that conditional moment inequalities point-identify θ if Θ_I is a singleton. Otherwise, it is said that conditional moment inequalities partially identify θ. For example, partial identification may happen when the conditional moment inequalities arise from a game-theoretic model with multiple equilibria. Moreover, the parameter θ may be weakly identified meaning that Θ_I is a singleton but information on Θ_I contained in the data is limited even in large samples. My method leads to robust tests with the correct asymptotic size regardless of whether the parameter θ is identified, weakly identified, and/or partially identified.

Testing moment inequalities has been a popular research topic in the econometrics literature recently; a nice review of this literature can be found in Canay and Shaikh (2016). The papers from this literature that are most closely related to mine are Andrews and Shi (2013), Chernozhukov et al. (2013b), Lee et al. (2013), Ponomareva (2010), Armstrong (2014a, 2015a), and Armstrong and Chan (2016). The method of Andrews and Shi (2013) is based on converting conditional moment inequalities 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 Armstrong (2015a) that their test has a relatively low rate of uniform consistency. The method of Chernozhukov et al. (2013b) is based on estimating moment functions nonparametrically. 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).[3,4]

The test of Ponomareva (2010) is related to that used in this paper but she uses only one value of the smoothing parameter, which makes her test non-adaptive. Lee et al. (2013) developed a test based on the minimum distance statistic in the one-sided L_p-norm and kernel estimates of moment functions. The advantage of their approach comes from simplicity of their critical value for the test, which is an appropriate quantile of the standard Gaussian distribution. Their test is not adaptive, however, since only one bandwidth value is used. Armstrong (2015a) developed a new method for computing the critical value for the test statistic of Andrews and Shi (2013) 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 identification. Armstrong (2014a), 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. After this paper had been made publicly available, Armstrong and Chan (2016) made an important contribution by constructing a test based on a statistic 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 they provided an explicit form of the limit distribution that was previously unknown.

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

The rest of the paper is organized as follows. Section 4 formally introduces the tests. The main results of the paper are presented in Section 3. A Monte Carlo simulation study is described in Section 4. 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%. Section 5 concludes. Finally, implementation details and all proofs are contained in the Appendix.

2. TESTS

In this section, I develop two tests of the null hypothesis $H_0: \theta = \theta_0$ against the alternative $H_a: \theta \neq \theta_0$ in the model (1). The tests are based on the same statistic but differ in critical values. Both tests reject H_0 if the statistic exceeds the corresponding critical value.

For $j = 1, \ldots, p$, denote $f_j(X) = \mathrm{E}[m_j(X, W, \theta_0) \mid X]$ and $\varepsilon_j = m_j(X, W, \theta_0) - f_j(X)$. Let $f(X) = (f_1(X), \ldots, f_p(X))^T$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)^T$. Also, let $\Sigma(X) = \mathrm{E}[\varepsilon\varepsilon^T \mid X]$ and for $i = 1, \ldots, n$, let $\Sigma_i = \Sigma(X_i)$. Section 2.1 defines the test statistic. Section 2.2 gives two bootstrap methods to simulate the critical values. The first method is based on a one-step procedure and is referred to as the plugin method. The second method employs a refined moment selection procedure and is referred to as the RMS method. Calculating both the statistic and the critical values requires having estimates of Σ_i's. Section 2.3 shows how to construct these estimates.

2.1. Test Statistic

Observe that under H_0, we have

$$f_j(X) \leqslant 0, \quad \text{for all } j = 1, \ldots, p,$$

where the inequalities hold almost surely. Therefore, one can form a test of H_0 using estimators $\widehat{f_j}$ of the functions f_j. In this paper, I employ kernel estimators. For $i = 1,\dots,n$ and $j = 1,\dots,p$, denote $Y_{i,j} = m_j(X_i, W_i, \theta_0)$, and let $Y_i = (Y_{i,1},\dots,Y_{i,p})^T$. Also, let $\varepsilon_{i,j} = Y_{i,j} - f_j(X_i)$ and $\varepsilon_i = (\varepsilon_{i,1},\dots,\varepsilon_{i,p})^T$. Let $K\colon \mathbb{R}^d \to \mathbb{R}_+$ be a positive kernel function, that is, any function on $\mathbb{R}^d$ that takes positive values and integrates to one. For all bandwidth values $h > 0$, let $K_h(x) = K(x/h)/h^d$. For $x', x'' \in \mathbb{R}^d$, define the weight function

$$w_h(x', x'') = \frac{K_h(x' - x'')}{\sum_{i=1}^n K_h(x' - X_i)}.$$

Then for x in the support of X, the kernel estimator of $f_j(x)$ is

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

Conditional on $X_1^n = (X_i)_{i=1}^n$, the variance of $\widehat{f}_{j,h}(x)$ is

$$V_{j,h}^2(x) = \sum_{i=1}^n w_h^2(x, X_i) \Sigma_{i,jj},$$

where $\Sigma_{i,j_1 j_2}$ denotes the (j_1, j_2) component of $\Sigma_i = \Sigma(X_i)$. Also, for $i = 1,\dots,n$, let $\widehat{\Sigma}_i$ be an estimator of Σ_i. Then the estimator of $V_{j,h}^2(x)$ is

$$\widehat{V}_{j,h}^2(x) = \sum_{i=1}^n w_h(x, X_i) \widehat{\Sigma}_{i,jj},$$

where $\widehat{\Sigma}_{i,j_1 j_2}$ denotes the (j_1, j_2) component of $\widehat{\Sigma}_i$.

Next, consider a finite set of bandwidth values $\mathcal{H} = \mathcal{H}_n = \{h = h_{\max} a^k : h \geqslant h_{\min}, k = 0, 1, 2, \dots\}$ for some $h_{\max} = h_{\max,n}$, $h_{\min} = h_{\min,n}$ where $h_{\max} > h_{\min}$, and $a \in (0,1)$. For the asymptotic analysis in this paper, I will assume that as the sample size n increases, the largest bandwidth value, $h_{\max}$, remains bounded below from zero, and the smallest bandwidth value, $h_{\min}$, converges to zero. Practical recommendations on how to choose different tuning parameters for the tests are given in Appendix A.

Finally, let $\mathcal{S} = \mathcal{S}_n = \{(i, j, h)\colon 1 \leqslant i \leqslant n; 1 \leqslant j \leqslant p; h \in \mathcal{H}\}$. Using this notation, I define the test statistic as

$$T = \max_{(i,j,h)\in\mathcal{S}} \frac{\widehat{f}_{j,h}(X_i)}{\widehat{V}_{j,h}(X_i)}. \tag{2}$$

Thus, the test statistic is based on the studentized kernel estimates of the functions f_j at the points $(X_i)_{i=1}^n$ corresponding to many different bandwidth values $h \in \mathcal{H}$.[5]

Let me now explain why using many different bandwidth values $h \in \mathcal{H}$ instead of just one value is useful. Without loss of generality, assume that the (global) maximum in (2) is achieved at $j = 1$. Suppose that the function f_1 is nearly flat in the neighborhood of its maximum. Then f_1 is positive on a large subset of its domain whenever its maximal value is strictly 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 has a sharp peak at the maximum, there may not exist a large subset of its domain where it is positive. Hence, large bandwidth values may not yield large values of T, in which case the maximum will be achieved by a small bandwidth value. Since ex ante the shape of f_1 in the neighborhood of its maximum is unknown, it is beneficial to use many different bandwidth values jointly, and let the data determine the best value. Thus, the test statistic T adapts to the smoothness level of the functions f_j leading to tests with sound power properties.

Another important feature of the test statistic T is that it uses *local* optimal bandwidth values, that is, for each i and j, the test looks for an optimal bandwidth value separately. This is in contrast with *global* optimal bandwidth values that could come, for example, from cross-validation. The use of local optimal bandwidth values is important when the smoothness of the functions $x \mapsto f_j(x)$ varies over x and j.

2.2. Critical Values

This subsection explains how to simulate a critical value $c_{1-\alpha}$ for the statistic T to obtain a test with asymptotic size $\alpha \in (0, 1/2)$ using two different bootstrap methods. The first method is based on a one-step procedure and gives the plugin critical value $c_{1-\alpha}^{PIA}$. The second method employs a refined moment selection procedure and gives the RMS critical value $c_{1-\alpha}^{RMS}$. The plugin test rejects H_0 if $T > c_{1-\alpha}^{PIA}$, and the RMS test rejects H_0 if $T > c_{1-\alpha}^{RMS}$.

The first method relies on two observations. First, one can show that in asymptotics, under H_0, the distribution of T is first order stochastically dominated by the distribution of T corresponding to the model with $f_j \equiv 0$ for all $j = 1, \dots, p$. Second, results in Chernozhukov, Chetverikov, and Kato (2013a), Chernozhukov, Chetverikov, and Kato (2014a), and Chernozhukov, Chetverikov, and Kato (2014b) imply that the conditional distribution of T given $(X_i)_{i=1}^n$ asymptotically depends on the distribution of ε_i's only via Σ_i's. These observations suggest that one can simulate a critical value for the test, $c_{1-\alpha}^{PIA}$, using the following procedure:

(1) For each $i = 1, \dots, n$, simulate $\widetilde{Y}_i = (\widetilde{Y}_{i,1}, \dots, \widetilde{Y}_{i,p})^T \sim N(0_p, \widehat{\Sigma}_i)$ independently across i.

(2) Calculate $T^{PIA} = \max_{(i,j,h)\in S} \sum_{l=1}^n w_h(X_i, X_l)\widetilde{Y}_{l,j}/\widehat{V}_{j,h}(X_i)$.

(3) Repeat steps 1 and 2 independently B times for some large B to obtain $\left(T_b^{PIA}\right)_{b=1}^B$.

(4) Let $c_{1-\alpha}^{PIA}$ be the $(1-\alpha)$ empirical quantile of $\left(T_b^{PIA}\right)_{b=1}^B$.

Here 0_p denotes the vector in $\mathbb{R}^p$ whose all elements are zero.

The second method is based on the refined moment selection procedure. It gives a more powerful test without sacrificing asymptotic size of the test. The method partially fixes the slackness arising from replacing the distribution of T under H_0 by the distribution of T corresponding to the model with $f_j \equiv 0$ for all $j = 1, \ldots, p$. To describe the method, let $\gamma = \gamma_n < \alpha/2$ be some small positive truncation parameter. For the asymptotic analysis in this paper, I will assume that γ converges to zero as the sample size n increases. See Appendix A for practical recommendations on how to choose γ. In addition, let $c_{1-\gamma}^{PIA}$ be the plugin critical value corresponding to the level γ. Denote

$$\mathcal{S}^{RMS} = \mathcal{S}_n^{RMS} = \left\{ (i,j,h) \in \mathcal{S} \colon \frac{\widehat{f}_{j,h}(X_i)}{\widehat{V}_{j,h}(X_i)} > -2c_{1-\gamma}^{PIA} \right\}.$$

Then the RMS critical value $c_{1-\alpha}^{RMS}$ is given by the following procedure:

(1) For each $i = 1, \ldots, n$, simulate $\widetilde{Y}_i = (\widetilde{Y}_{i,1}, \ldots, \widetilde{Y}_{i,p})^T \sim N(0_p, \widehat{\Sigma}_i)$ independently across i.

(2) Calculate $T^{RMS} = \max_{(i,j,h) \in \mathcal{S}^{RMS}} \sum_{l=1}^n w_h(X_i, X_l) \widetilde{Y}_{l,j} / \widehat{V}_{j,h}(X_i)$.

(3) Repeat steps 1 and 2 independently B times for some large B to obtain $\left(T_b^{RMS}\right)_{b=1}^B$.

(4) Let $c_{1-\alpha}^{RMS}$ be the $(1-\alpha)$ empirical quantile of $\left(T_b^{RMS}\right)_{b=1}^B$.

Validity of both $c_{1-\alpha}^{PIA}$ and $c_{1-\alpha}^{RMS}$ in terms of yielding the tests with asymptotic size control will be established below in Section 3.

2.3. Estimating Σ_i's

Let me now explain how one can estimate Σ_i's. There are many methods in the literature that allow to estimate Σ_i's; for scalar-valued Y_i's, available estimators are described in Horowitz and Spokoiny (2001). All those estimators can be immediately generalized to vector-valued Y_i's. For concreteness, I describe one estimator here. Choose a bandwidth value $b = b_n > 0$ with $b \to 0$ as $n \to \infty$. For $i = 1, \ldots, n$, let $J(i) = \{ j = 1, \ldots, n \colon \|X_j - X_i\| < b\}$ denote all observations X_j that are in b-neighborhood of X_i. If $J(i)$ has an odd number of elements, drop one arbitrarily selected observation from this set. Partition $J(i)$ into pairs using a map $\mathcal{K} \colon J(i) \to J(i)$ satisfying $\mathcal{K}(j) \neq j$ and $\mathcal{K}(\mathcal{K}(j)) = j$ for all $j \in J(i)$, so that each j is paired with $\mathcal{K}(j)$. Let $|J(i)|$ denote the number of elements in $J(i)$. Then for all $i = 1, \ldots, n$, one can estimate Σ_i by

$$\widehat{\Sigma}_i = \frac{1}{2|J(i)|} \sum_{j \in J(i)} (Y_{\mathcal{K}(j)} - Y_j)(Y_{\mathcal{K}(j)} - Y_j)^T.$$

Lemma B2 in the Appendix gives certain conditions that ensure that this estimator will be consistent for Σ_i uniformly over $i = 1, \ldots, n$ with a polynomial rate of convergence as required in A5 below. To choose the bandwidth value b in practice, one can use some version of cross validation.

Intuition behind this estimator is as follows. Note that $\mathcal{K}(j)$ is chosen so that $X_{\mathcal{K}(j)}$ is close to X_j. If the functions f_j are continuous,

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

so that

$$\mathrm{E}\big[(Y_{\mathcal{K}(j)} - Y_j)(Y_{\mathcal{K}(j)} - Y_j)^T \mid (X_i)_{i=1}^n\big] \approx \Sigma_{\mathcal{K}(j)} + \Sigma_j$$

since $\varepsilon_{\mathcal{K}(j)}$ is independent of ε_j. If b is small enough and $\Sigma(\cdot)$ is continuous, then $\Sigma_{\mathcal{K}(j)} + \Sigma_j \approx 2\Sigma_i$ since $\|X_{\mathcal{K}(j)} - X_i\| \leqslant b$ and $\|X_j - X_i\| \leqslant b$, and so $\widehat{\Sigma}_i$'s should be close to Σ_i's by the law of large numbers.

3. MAIN RESULTS

This section presents main results of the paper. Section 3.1 gives regularity conditions. Section 3.2 describes size properties of the tests. Section 3.3 explains the behavior of the tests under a fixed alternative and establishes consistency of the tests. Section 3.4 derives the rate of consistency of the tests against local one-dimensional alternatives. Section 3.5 derives the rate of uniform consistency against certain classes of smooth alternatives. Section 3.6 presents the minimax rate-optimality result.

3.1. Assumptions

Let c_1, c_h, C_1, q, τ_0, τ, and L be strictly positive and finite constants independent of the sample size n, where $q > 4$, $\tau_0 > d/(q-2)$, and $0 < \tau \leqslant 1$. Also, for $x \in \mathbb{R}^d$ and $\epsilon > 0$, let $B(x,\epsilon)$ denote a ball in $\mathbb{R}^d$ with center at x and radius ϵ, that is, $B(x,\epsilon) = \{y \in \mathbb{R}^d : \|y - x\| < \epsilon\}$. Moreover, let λ denote the Lebesgue measure on $\mathbb{R}^d$. In addition, for a set $\mathcal{A}$ in $\mathbb{R}^d$, let $\operatorname{diam}(\mathcal{A})$ denote the diameter of $\mathcal{A}$, that is, $\operatorname{diam}(\mathcal{A}) = \sup_{x',x'' \in \mathcal{A}} \|x' - x''\|$. Finally, for a $p \times p$ matrix A, let $\|A\|$ denote the spectral norm of A corresponding to the Euclidean norm on $\mathbb{R}^p$, that is,

$$\|A\| = \sup_{x \in \mathbb{R}^p:\ \|x\|=1} \|Ax\|. \tag{3}$$

Results in this paper will be proven under the following assumptions.

A1. (i) The support $\mathcal{X}$ of X is a closed set in $\mathbb{R}^d$ with the properties that $\operatorname{diam}(\mathcal{X}) \leqslant C_1$ and that $\lambda(B(x,\epsilon) \cap \mathcal{X}) \geqslant c_1 \lambda(B(x,\epsilon))$ for all $x \in \mathcal{X}$ and $\epsilon \in (0,1)$. (ii) The pdf of X is bounded from above by C_1 and from below by c_1 on $\mathcal{X}$.

This is a mild assumption on the distribution of X. The first part of A1(i) requires that X has a bounded support. The second part of A1(i) is satisfied, for example, if $\mathcal{X}$ is convex. A1(ii) imposes the condition that X has a continuous distribution with the pdf bounded from above and away from zero on its support.[6]

A2. For all $x \in \mathcal{X}$ and $j = 1, \ldots, p$, the random vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)^T$ satisfies $\mathrm{E}[|\varepsilon_j|^q \mid X = x] \leqslant C_1^q$ and $\mathrm{E}[\varepsilon_j^2 \mid X = x] \geqslant c_1^2$.

This is a mild assumption on the conditional distribution of ε given X. I assume that the variance of ε_j's 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 main ideas.

A3. The maximal and minimal bandwidth values $h_{\max}$ and $h_{\min}$ satisfy $h_{\max} = \max_{1 \leqslant i,l \leqslant n} \|X_i - X_l\|$ and $h_{\min} = c_h^{1/d} h_{\max}(\log n/n)^{1/(2\tau_0+d)}$.

According to this assumption, the maximal bandwidth value, $h_{\max}$, is chosen to match the diameter of the set of design points $(X_i)_{i=1}^n$. 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. 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 $h_{\max} \to 0$, the tests will have relatively low power in detecting global deviations from H_0 where one of the functions f_j is positive on a large set.

A4. (i) The kernel K is positive and supported on $\{x \in \mathbb{R}^d : \|x\| \leqslant 1\}$. (ii) For all $x \in \mathbb{R}^d$, the kernel K satisfies $K(x) \leqslant C_1$ and for all $x \in \mathbb{R}^d$ with $\|x\| \leqslant 1/2$, it satisfies $K(x) \geqslant c_1$.

Many kernels satisfy this assumption. For example, one can use rectangular, triangular, parabolic, or biweight kernels; see Tsybakov (2009). On the other hand, the requirement that the kernel is positive excludes higher-order kernels, which could be used to estimate the functions f_j more effectively if these functions are sufficiently smooth. I require positive kernels because they have an invariance property 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 T might be caused by large values of the bias term relative to the standard deviation of the estimator even when all functions f_j are negative. However, to achieve under-smoothing, one has to know the smoothness properties of f_j's. 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.

A5. The estimators $\widehat{\Sigma}_i$ satisfy $\mathrm{P}(\max_{i=1,\ldots,n} \|\widehat{\Sigma}_i - \Sigma_i\| > C_1 n^{-c_1}) \leqslant C_1 n^{-c_1}$.

This assumption is satisfied for $\widehat{\Sigma}_i$'s described in Section 2.3 under mild regularity conditions; see Lemma B2 in the Appendix.

A6. The truncation parameter γ satisfies $\gamma = \gamma_n \leqslant C_1 n^{-c_1}$.

This assumption is easy to satisfy by selecting γ appropriately; see Appendix A for practical recommendations.

Recall that $L > 0$ and $0 < \tau \leqslant 1$. Let $\mathcal{F}(\tau, L)$ be the set of all functions $g : \mathcal{X} \to \mathbb{R}$ satisfying

$$|g(x') - g(x'')| \leqslant L\|x' - x''\|^{\tau}$$

for all $x', x'' \in \mathcal{X}$. In the literature, $\mathcal{F}(\tau, L)$ is typically referred to as a Hölder ball.

A7. For all $j = 1, \ldots, p$, the functions f_j satisfy $f_j \in \mathcal{F}(\tau, L)$.

For simplicity of notation, I assume that all functions f_j have the same smoothness properties in the sense that they belong to the same Hölder ball $\mathcal{F}(\tau, L)$. Note that Hölder balls $\mathcal{F}(\tau, L)$ can also be defined for $\tau > 1$ but in this paper, I only consider the case $\tau \leqslant 1$, although the result on the rate of uniform consistency against alternatives in the ball $\mathcal{F}(\tau, L)$, presented in Theorem 4, can be extended to the case $\tau \in (1, 2]$ in the expense of some technicalities. The result can not be extended to the case $\tau > 2$ since I rely on positive kernels, which exclude higher-order kernels that are needed to effectively estimate functions with Hölder smoothness $\tau > 2$ and achieve the optimal rate of uniform consistency against Hölder balls $\mathcal{F}(\tau, L)$ with $\tau > 2$.

3.2. Size Properties of the Test

Analysis of size properties of the tests developed in this paper is complicated because it is unknown whether the test statistic T has a limit distribution.[7] Instead, I use a finite sample method developed in Chernozhukov et al. (2013a), Chernozhukov et al. (2014a), and Chernozhukov et al. (2014b). For each sample size n, this method gives an upper bound on the uniform distance between the cdf of the test statistic T and the cdf the test statistic T would have if the random vectors $(\varepsilon_i)_{i=1}^n$ were Gaussian. My first theorem in this paper states that the plugin and RMS tests have correct asymptotic size, and the tests are nonconservative as the size of the tests converges to the nominal level α as $n \to \infty$. All proofs are given in the Appendix.

THEOREM 1 (Asymptotic size control). *Let $P = PIA$ or RMS. Suppose that A1–A6 hold. In addition, suppose that H_0 holds. Then there exist constants $c, C > 0$ that depend only on c_1, c_h, C_1, a, p, d, q, α, and τ_0 such that*

$$\mathrm{P}\big(T > c^P_{1-\alpha}\big) \leqslant \alpha + Cn^{-c}. \tag{4}$$

If, in addition, $f_j(x) = 0$ for all $x \in \mathcal{X}$ and $j = 1, \ldots, p$, then

$$\big|\mathrm{P}\big(T > c^P_{1-\alpha}\big) - \alpha\big| \leqslant Cn^{-c}. \tag{5}$$

Moreover, as long as $\gamma = \gamma_n$ is such that $\log(1/\gamma) \leqslant C_1 \log n$, if for some $\mathcal{J} \subset \{1, \ldots, p\}$, it follows that $f_j(x) = 0$ for all $x \in \mathcal{X}$ and $j \in \mathcal{J}$ but $f_j(x) \leqslant -c_1$

for all $x \in \mathcal{X}$ and $j \notin \mathcal{J}$, then

$$\left|\mathrm{P}\left(T > c^{RMS}_{1-\alpha}\right) - \alpha\right| \leqslant Cn^{-c}. \tag{6}$$

Comment 1 (Advantages of the theorem, I)**.** One of the advantages of this theorem is that it shows that the probability of rejecting H_0 when H_0 holds can exceed the nominal level α only by a *polynomially small* (in n) number Cn^{-c}. This implies that the bootstrap procedures developed in this paper provide high quality inference in finite samples. This contribution is important since asymptotic approximations of suprema of processes, as in the test statistic T here, typically provide only *logarithmically* (in n) small approximation error; see, for example, Hall (1991).

Comment 2 (Advantages of the theorem, II)**.** The theorem provides bounds on the difference between the probability of rejecting H_0 when H_0 holds and the nominal level α that depend only on the constants c_1, c_h, C_1, a, p, d, q, α, and τ_0. Therefore, the bounds are the same for all data-generating processes satisfying the assumptions with the same constants, and so the bound holds *uniformly* over all these data-generating processes. For example (4) implies that

$$\sup_F \mathrm{P}_F\left(T > c^{P}_{1-\alpha}\right) \leqslant \alpha + Cn^{-c},$$

where F denotes the distribution of the pair (X, W), and the supremum is taken over all F that satisfy Assumptions A1–A6 with the same constants c_1, c_h, C_1, a, p, d, q, α, and τ_0. This also serves as a guarantee that the test controls size well in finite samples.

Comment 3 (Comparison of regularity conditions with those in Andrews and Shi (2013))**.** Although the results in Theorem 1 do not require the functions f_j to be continuous (A7 is not imposed), the results rely upon A5, and constructing estimators $\widehat{\Sigma}_i$ that satisfy A5 typically requires the functions f_j to be smooth; see Lemma B2 in the Appendix. In contrast, the results of Andrews and Shi (2013) do not require such smoothness conditions, which is one of the advantages of their tests. Another advantage of their tests is that they only require $\mathrm{E}[|m_j(X, W, \theta_0)|^{2+\delta}]$ to be finite for all $j = 1, \ldots, p$ and some $\delta > 0$, whereas I require $q > 4$ finite conditional moments of $\varepsilon_j = m_j(X, W, \theta_0) - \mathrm{E}[m_j(X, W, \theta_0) \mid X]$ given X; see A2. Finally, Andrews and Shi (2013) do not impose the condition that the conditional variance of ε_j's given X is bounded below from zero, as I do here; see A2.

Comment 4 (Other choices of test statistic)**.** Inspecting the proof of Theorem 1 reveals that, in fact, one does not have to use weights $w_h(x, X_i)$ based on kernel estimators to form a test statistic T. Alternatively, one can consider any test statistic of the form

$$T' = \max_{l=1,\ldots,L} \max_{j=1,\ldots,p} \frac{\sum_{i=1}^{n} w_l(X_i) Y_{i,j}}{\sqrt{\sum_{i=1}^{n} w_l^2(X_i) \widehat{\Sigma}_{i,jj}}},$$

where $(w_l)_{l=1}^L$ is some sequence of weight functions mapping $\mathcal{X}$ into $\mathbb{R}_+ = \{y \in \mathbb{R}\colon y \geqslant 0\}$ and $L = L_n$ is some positive integer, possibly depending on n. The results of Theorem 1 apply as long as $\log L \leqslant C_1 \log n$ and these weight functions depend on the data only via $(X_i)_{i=1}^n$ and satisfy the following condition: with probability at least $1 - C_1 n^{-c_1}$,

$$\frac{w_l^2(X_i)\Sigma_{i,jj}}{\sum_{k=1}^n w_l^2(X_k)\Sigma_{k,jj}} \leqslant C_1 n^{-2/q-c_1}$$

for all $i = 1,\ldots,n$, $j = 1,\ldots,p$, and $l = 1,\ldots,L$. This is the condition required to apply Lemma B10 in the proof of Theorem 1. Under A1 and A2, this condition can be verified, for example, if

$$w_l(X_i) = \sum_{r=1}^R K\left(\frac{x_r - X_i}{h}\right),$$

where R is some positive integer, $l = (x_1,\ldots,x_R,h)$, $\{x_1,\ldots,x_R\} \subset \mathcal{X}$, and $h = h_n$ satisfies $(nh^d)^{-1} \leqslant C_1 n^{-2/q-c_1}$. The test statistic T' based on such weight functions would have high power against alternatives with many small peaks.

Comment 5 (Choice of γ)**.** One of the conditions of Theorem 1, A6, requires that γ satisfies $\gamma = \gamma_n \leqslant C_1 n^{-c_1}$. The proof of Theorem 1 reveals, however, that even if this condition is violated, the result (4) for $P = RMS$ still holds as long as one does a finite-sample adjustment and defines $c_{1-\alpha}^{RMS}$ as the $(1-\alpha+2\gamma)$ empirical quantile of $(T_b^{RMS})_{b=1}^B$, instead of taking the $(1-\alpha)$ empirical quantile of $(T_b^{RMS})_{b=1}^B$; see Section 2.2. A related adjustment has been independently developed in Romano, Shaikh, and Wolf (2014), although they derived a somewhat different moment selection procedure. This type of adjustment can be traced back at least to Loh (1985). In addition, if instead of assuming that $\gamma = \gamma_n \leqslant C_1 n^{-c_1}$, one imposes the condition that $\gamma = \gamma_n \to 0$, the quantities Cn^{-c} in (4) and (5) when $P = RMS$ would be replaced by $o(1)$. No finite-sample adjustment of $c_{1-\alpha}^{RMS}$ is required in this case.

3.3. Consistency Against a Fixed Alternative

In this section, I show that the tests developed in this paper are consistent against any fixed alternative outside of the identified set, that is, I show that if $\theta_0 \notin \Theta_I$ and θ_0 does not depend on n, the tests will reject H_0 with probability approaching one.

THEOREM 2 (Consistency against fixed alternatives). *Let $P = PIA$ or RMS. Suppose that A1–A6 hold. In addition, suppose that the functions f_j are continuous for all $j = 1,\ldots,p$. Moreover, suppose that $\theta_0 \notin \Theta_I$. Then*

$$\mathrm{P}\left(T > c_{1-\alpha}^P\right) \to 1.$$

3.4. Consistency Against Local One-Dimensional Alternatives

In this section, I derive the rate of consistency of the tests developed in this paper against local one-dimensional alternatives. I will assume that the tested parameter value, θ_0, is now allowed to depend on n, that is, $\theta_0 = \theta_{0,n}$. Consequently, the functions f_j are also allowed to depend on n, that is, $f_j = f_{j,n}$.

For $j = 1, \ldots, p$, let $f_j^0 \colon \mathcal{X} \to \mathbb{R}$ be functions such that

$$\rho_0 = \sup_{x \in \mathcal{X}} \max_{j=1,\ldots,p} f_j^0(x) > 0, \tag{7}$$

and let $(a_n)_{n \geqslant 1}$ be a sequence of positive numbers converging to zero. I will assume that $\theta_{0,n}$ is such that

$$f_{j,n} = a_n f_j^0, \quad \text{for all } j = 1, \ldots, n \text{ and } n \geqslant 1. \tag{8}$$

As long as $a_n > 0$, it follows that $\theta_{0,n} \notin \Theta_I$. I refer to the sequence $(\theta_{0,n})_{n \geqslant 1}$ satisfying (8) with $a_n > 0$ for all $n \geqslant 1$ as local one-dimensional alternatives. In the theorem below, I show that the tests developed in this paper are consistent against such alternatives if $a_n (n/\log n)^{1/2} \to \infty$.

THEOREM 3 (Consistency against local one-dimensional alternatives). *Let $P = PIA$ or RMS. Suppose that $\theta_0 = \theta_{0,n}$ is such that A1–A6 hold for all $n \geqslant 1$ and, in addition, the functions $f_j = f_{j,n}$ are given by* (8) *for some continuous functions f_j^0 satisfying* (7). *Moreover, suppose that $a_n (n/\log n)^{1/2} \to \infty$. Then*

$$\mathrm{P}\big(T > c_{1-\alpha}^P\big) \to 1.$$

3.5. Uniform Consistency Against Hölder Smoothness Classes

In this section, I derive the rate of uniform consistency of the tests developed in this paper against classes $\mathcal{F}(\tau, L)$. As in Section 2, for a given pair of random vectors (X, W) and given parameter value θ_0, define $f_j(X) = \mathrm{E}[m_j(X, W, \theta_0) \mid X]$ for all $j = 1, \ldots, p$ and set $f(X) = (f_1(X), \ldots, X_p(X))^T$. Then one can measure the distance between θ_0 and the identified set Θ_I by

$$\rho(\theta_0, \Theta_I) = \sup_{x \in \mathcal{X}} \max_{j=1,\ldots,p} \max(f_j(x), 0).$$

It follows that $\theta_0 \notin \Theta_I$ if and only if $\rho(\theta_0, \Theta_I) > 0$. Let $(a_n)_{n \geqslant 1}$ be a sequence of positive numbers converging to zero, and let Θ_{a_n} be a set of all $\theta_0 \in \Theta$ such that $\rho(\theta_0, \Theta_I) \geqslant a_n$. In this section, it will be convenient to make dependence of T and $c_{1-\alpha}^P$, $P = PIA$ or RMS, on θ_0 explicit. Therefore, I will write $T(\theta_0)$ and $c_{1-\alpha}^P(\theta_0)$ instead of T and $c_{1-\alpha}^P$, respectively.

THEOREM 4 (Uniform consistency against Hölder smoothness classes). *Let $P = PIA$ or RMS. Suppose that A1–A7 hold for all $\theta_0 \in \Theta$. In addition, suppose* *that $\tau_0 < \tau \leqslant 1$. Moreover, suppose that $a_n(n/\log n)^{\tau/(2\tau+d)} \to \infty$. Then*

$$\inf_{\theta_0 \in \Theta_{a_n}} \mathrm{P}\big(T(\theta_0) > c^P_{1-\alpha}(\theta_0)\big) \to 1. \tag{9}$$

3.6. Lower Bound on the Minimax Rate of Testing

In this section, I derive a lower bound on the minimax rate of testing. Let $\mathcal{E}$ denote the set of all random vectors $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)^T$ that satisfy A2 and are such that $\mathrm{E}[\varepsilon \mid X] = 0_p$. Also, let $\times_{j=1}^p \mathcal{F}(\tau, L)$ denote the set of all vector-valued functions $f = (f_1, \ldots, f_p)^T$ such that $f_j \in \mathcal{F}(\tau, L)$ for all $j = 1, \ldots, p$, where $\tau \in (0,1]$ and $L > 0$. In addition, let $(a_n)_{n \geqslant 1}$ be a sequence of positive numbers converging to zero, and let $(\times_{j=1}^p \mathcal{F}(\tau, L))_{a_n}$ be a set of all $f \in \times_{j=1}^p \mathcal{F}(\tau, L)$ such that

$$\sup_{x \in \mathcal{X}} \max_{j=1,\ldots,p} f_j(x) \geqslant a_n.$$

Let $\varepsilon \in \mathcal{E}$ and $f \in \times_{j=1}^p \mathcal{F}(\tau, L)$. Also, let $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ be a random sample from the distribution of (X, Y), where $Y = f(X) + \varepsilon$. I consider tests that reject H_0 with probability $\phi_n(\mathcal{D}_n)$ for some function ϕ_n. I use $\mathrm{E}_{f,\varepsilon}[\phi_n(\mathcal{D}_n)]$ to denote the expectation with respect to $\mathcal{D}_n$ with indices f and ε emphasizing the dependence of the distribution of $\mathcal{D}_n$ on f and ε. The following theorem derives a lower bound on the minimax rate of testing.

THEOREM 5 (Lower bound). *Suppose that A1 holds. Consider any sequence of tests $(\phi_n)_{n \geqslant 1}$. Suppose that*

$$\sup_{\varepsilon \in \mathcal{E}} \mathrm{E}_{f,\varepsilon}[\phi_n(\mathcal{D}_n)] \leqslant \alpha + o(1)$$

as long as $f_j(x) = 0$ for all $x \in \mathcal{X}$ and $j = 1, \ldots, p$. In addition, suppose that $a_n(n/\log n)^{\tau/(2\tau+d)} \to 0$. Then

$$\inf_{f \in (\times_{j=1}^p \mathcal{F}(\tau, L))_{a_n}} \inf_{\varepsilon \in \mathcal{E}} \mathrm{E}_{f,\varepsilon}[\phi_n(\mathcal{D}_n)] \leqslant \alpha + o(1).$$

Comment 6 (On optimality of proposed tests)**.** This theorem shows that for any $\tau \in (0,1]$ and $L > 0$, no test that has asymptotic size control can have non-trivial power uniformly against alternatives belonging to the Hölder ball $\mathcal{F}(\tau, L)$ and having distance from the null hypothesis at least a_n as long as $a_n(n/\log n)^{\tau/(2\tau+d)} \to 0$. On the other hand, Theorem 4 shows that for any $L > 0$ and $\tau \in (\tau_0, 1]$, the tests developed in this paper are uniformly consistent against alternatives belonging to the Hölder ball $\mathcal{F}(\tau, L)$ and having distance from the null hypothesis at least a_n as long as $a_n(n/\log n)^{\tau/(2\tau+d)} \to \infty$. This implies that the tests developed in this paper are minimax rate-optimal (have optimal rate of uniform consistency) against Hölder balls $\mathcal{F}(\tau, L)$ for all $\tau \in (\tau_0, 1]$. Also, since τ does not have to be known to achieve the optimal rate, the tests are said to be adaptive with respect to the smoothness level τ.

Comment 7 (Relation of Theorem 5 to the literature)**.** The argument used in the proof of Theorem 5 is essentially due to Dumbgen and Spokoiny (2001). However, when two-sided alternatives are considered, that is, when the null hypothesis is that $f_j(x) = 0$ for all $x \in \mathcal{X}$ and $j = 1, \ldots, p$, and the alternative is that $\sup_{x\in\mathcal{X}} \max_{j=1,\ldots,p} |f_j(x)| > 0$, the same rate of uniform consistency as obtained in Theorem 5 has been first derived by Ingster (1987) for the related Gaussian white noise model. In the same setting as that in Ingster (1987), in addition to the optimal rate of uniform consistency, Lepski and Tsybakov (2000) also derived the exact separation constant, and thus fully characterized the set of alternatives against which non-trivial testing is possible from the minimax point of view. Guerre and Lavergne (2002) and Horowitz and Spokoiny (2001) derived minimax rate-optimal specification tests for the null hypothesis that the mean regression function belongs to some parametric class against general nonparametric alternatives. The review of the minimax testing literature can be found in Ingster and Suslina (2003). See also Armstrong (2015b) for some recent contributions on minimax testing with shape constraints.

Comment 8 (Comparison of testing power with that in Andrews and Shi (2013))**.** The tests of Andrews and Shi (2013) are consistent against local one-dimensional alternatives like those studied in Theorem 3 as long as $a_n n^{1/2} \to \infty$. In contrast, my tests are consistent against the same alternatives if $a_n(n/\log n)^{1/2} \to \infty$, and so my tests are expected to have lower power against these alternatives. On the other hand, it follows from results in Armstrong (2015a) that the tests of Andrews and Shi (2013) do not achieve the optimal rate of uniform consistency against Hölder balls $\mathcal{F}(\tau, L)$ for $\tau \in (0,1]$ and $L > 0$, whereas my tests do achieve the optimal rate against these alternatives as long as $\tau_0 < \tau \leqslant 1$. Thus, the additional $\log^{1/2} n$ factor in the rate of consistency against local one-dimensional alternatives that my tests have relative to those in Andrews and Shi (2013) is a cost for achieving the optimal rate of uniform consistency against these Hölder balls. A comprehensive comparison of different tests from the minimax optimality point of view can be found in Armstrong (2014b).

4. 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 tests accurately maintain size in finite samples. Second, I compare relative advantages and disadvantages of my tests and the tests of Andrews and Shi (2013), Chernozhukov et al. (2013b), and Lee et al. (2013). The methods of Andrews and Shi (2013) and Lee et al. (2013) are most appropriate for detecting flat alternatives, which represent local one-dimensional alternatives. These methods have low power against alternatives with peaks, however. The test of Chernozhukov et al. (2013b) 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 tests is their adaptiveness. In comparison with Andrews and Shi (2013) and Lee et al. (2013), my tests have higher power against alternatives with peaks. In comparison with Chernozhukov et al. (2013b), my tests have higher power when their rule-of-thumb techniques lead to an inappropriate bandwidth value.[8] For example, this happens when the underlying moment function is mostly flat but varies significantly in the region where the null hypothesis is violated (the case of spatially inhomogeneous alternatives; see Lepski and Spokoiny, 1999).

First Simulation Study. The data generating process is

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

where X_i's are equidistant on the $[-2,+2]$ interval,[9] 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: $L = 0.1$, $M = 0.2$, $m = 0.02$. Case 3: $L = M = 0$, $m = -0.02$. Case 4: $L = 2$, $M = 0.2$, $m = 0.2$. Note that $\mathrm{E}[Y_i \mid X_i] \leqslant 0$ for all $i = 1, \ldots, n$ in cases 1 and 2 while $\mathrm{E}[Y_i \mid X_i] > 0$ for some $i = 1, \ldots, n$ 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 level 10%. The results are based on 1,000 simulations for each specification.

For the test of Andrews and Shi (2013), 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

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

For the test of Chernozhukov et al. (2013b), 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 Chernozhukov et al. (2013b) and Lee et al. (2013) (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 non-adaptive. In particular, there is no formal theory on how to choose bandwidth values in their tests, so I follow their informal suggestions. For the test of Lee et al. (2013), I use their test statistic based on one-sided L_1-norm.

Parameters for the tests developed in this paper are chosen according to recommendations in Appendix A. Specifically, the smallest bandwidth value, $h_{\min}$, is set as $h_{\min} = 0.2 h_{\max} (\log n/n)^{1/3}$. The scaling parameter, a, equals 0.5 so that the set of bandwidth values is

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

I estimate Σ_i using the method of Rice (1984). Specifically, I rearrange the data so that $X_1 \leqslant \cdots \leqslant X_n$ 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 $\gamma = 0.046 \times n^{-0.17}$ to make meaningful comparisons with the test of Chernozhukov et al. (2013b), so that recommended γ for both papers coincide when $n = 250$ and $n = 500$. 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 tests are denoted as Adaptive test with plugin 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 Andrews and Shi (2013) and Lee et al. (2013) have highest rejection probabilities as expected from the theory. In this case, my test is less powerful in comparison with these tests and somewhat similar to the method of Chernozhukov et al. (2013b). This is compensated in case 4 where the null hypothesis is violated with the peak-shaped alternative. In this case, the power of my tests is much higher than that of competing tests. This is especially true for my test with the RMS critical value whose rejection

TABLE 1. Results of Monte Carlo experiments, $n = 250$

		Probability of rejecting null hypothesis						
Distribution ε	Case	AS, plugin	AS, GMS	LSW	CLR, V	CLR, $\widehat{V}$	Adaptive test, plugin	Adaptive test, RMS
Normal	1	0.096	0.091	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

Table 2. Results of Monte Carlo experiments, $n = 500$

		Probability of rejecting null hypothesis						
Distribution ε	Case	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	0.955	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

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 the RMS critical value has higher power than the test with the plugin critical value in case 4. So, among these two tests, I recommend the test with the RMS critical value. 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 Andrews and Shi (2013), I use their GMS critical value. The data-generating process is

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

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 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, $\mathrm{E}[Y_i \mid X_i] \leqslant 0$ for all $i = 1, \ldots, n$ 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 Andrews and Shi (2013). This figure shows that the rejection probability of the test developed in this paper is higher than that of the test of Andrews and Shi (2013) 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.

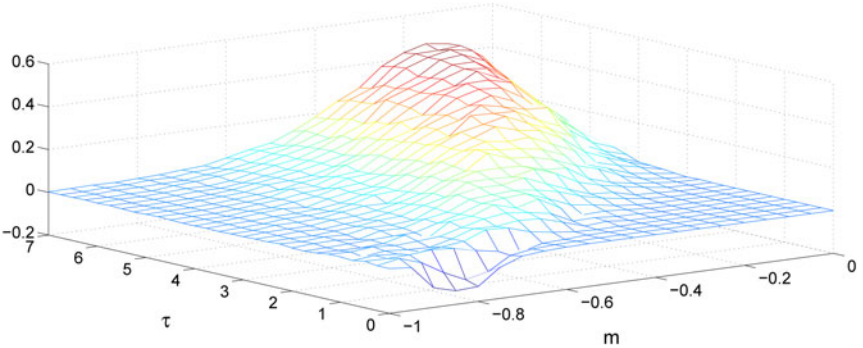

FIGURE 1. The difference between the rejection probabilities of the test developed in this paper and of the test of Andrews and Shi (2013) (with RMS and GMS critical values correspondingly). The nominal size is 10%. Results are based on 500 simulations. The figure shows that the rejection probability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

5. CONCLUSION

In this paper, I developed adaptive rate-optimal tests of parameter hypotheses in conditional moment inequality models. The tests are based on the studentized kernel estimates of the moment functions corresponding to many different bandwidth values and bootstrap critical values. The tests are computationally easy.

Regarding future work, I want to emphasize that although the tests presented in this paper have sound power properties against general nonparametric alternatives, given that the parameter of interest θ in the conditional moment inequalities model (1) is finite-dimensional, it may be possible to construct tests with even better power properties by directing the test statistic toward parametric alternatives implied by the finite-dimensional parameter θ.

NOTES

1. The preliminary results of the paper were first presented at the Econometric lunch at MIT on November 18, 2010. The slides of the presentation are available from the author upon request. The first public version of the paper appeared at arxiv.org in 2011.

2. Observe that the model (1) covers the models defined by conditional moment equalities as well because equalities can be represented by pairs of inequalities (in particular, doing so is compatible with all assumptions in the paper).

3. Efficiency of the test of Chernozhukov et al. (2013b) is achieved by using higher order kernel or series methods for estimating moment functions; both the tests of Andrews and Shi (2013) and the tests developed in this paper work with positive kernels, which exclude higher order kernels, and do not have this feature of the test of Chernozhukov et al. (2013b).

4. In the statistics literature, there recently have been developed techniques for adaptively selecting the appropriate smoothing parameter for tests like that in Chernozhukov et al. (2013b). 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, Gine and Nickl (2010). Deriving formal results on how the test of Chernozhukov et al. (2013b) works in combination with these adaptive smoothing parameter selection techniques would be an important direction for future research.

5. In principle, to form a test statistic, one could specify another grid of points at which the functions f_j 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 as n gets large, and in addition, this makes the statistic computationally simple.

6. In practice, X often contains discrete components. In this case, one can proceed as follows. Write $X = (X^{(d)}, X^{(c)})$ where $X^{(d)}$ and $X^{(c)}$ denote discrete and continuous components of X, respectively. Let $\mathcal{X}^{(d)}$ denote the support of $X^{(d)}$. For $\widetilde{j} = (j, x^d) \in \{1, \dots, p\} \times \mathcal{X}^d$, define $\widetilde{m}_{\widetilde{j}}(X^{(c)}, W, \theta) = m_j(X, W, \theta) \cdot I(X^{(d)} = x^d)$, where $I(\cdot)$ denotes the indicator function. Then inequalities (1) hold if and only if $\mathrm{E}\big[\widetilde{m}_{\widetilde{j}}(X^{(c)}, W, \theta) \mid X^{(c)}\big] \leqslant 0$ for all $\widetilde{j} = (j, x^d) \in \{1, \dots, p\} \times \mathcal{X}^d$. Since $X^{(c)}$ has a continuous distribution, to test H_0 against H_a, one can apply the tests developed in the previous section to these alternative (but equivalent) conditional moment inequalities.

7. After the first version of this paper appeared online, Armstrong and Chan (2016) derived the limit distribution of a closely related statistic in the case when $f_j(x) = 0$ for all $x \in \mathcal{X}$ and $j = 1, \dots, p$. By deriving the limit distribution, their paper solved a very difficult problem and made an important contribution to the literature, since their techniques can be applied in other related settings as well. My paper complements theirs as I show how to do bootstrap inference based on the statistic T.

8. When their rule-of-thumb works well and moment functions are sufficiently smooth, the test of Chernozhukov et al. (2013b) often yielded the best results in my simulations.

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

REFERENCES

Andrews, D.W.K. & X. Shi (2013) Inference based on conditional moment inequalities. *Econometrica* 81, 609–666.

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

Armstrong, T. (2014a) Weighted KS statistics for inference on conditional moment inequalities. *Journal of Econometrics* 181, 92–116.

Armstrong, T. (2014b) A note on minimax testing and confidence intervals in moment inequality models. Unpublished manuscript, arxiv:1412.5656.

Armstrong, T. (2015a) Asymptotically exact inference in conditional moment inequalities models. *Journal of Econometrics* 186, 51–65.

Armstrong, T. (2015b) Adaptive testing on a regression function at a point. *The Annals of Statistics* 43, 2086–2101.

Armstrong, T. & H. Chan (2016) Multiscale adaptive inference on conditional moment inequalities. *Journal of Econometrics* 194, 24–43.

Bhatia, R. (1997) *Matrix Analysis.* Springer.

Boucheron, S., G. Lugosi, & P. Massart (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press.

Canay, I. & A. Shaikh (2016) Practical and theoretical advances in inference for partially identified models. *Journal of Econometrics* 156, 408–425.

Chernozhukov, V., D. Chetverikov, & K. Kato (2013a) Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics* 41, 2786–2819.

Chernozhukov, V., D. Chetverikov, & K. Kato (2014a) Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields* 162, 47–70.

Chernozhukov, V., D. Chetverikov, & K. Kato (2014b) Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, arxiv:1412.3661.

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

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

Dumbgen, L. & V. Spokoiny (2001) Multiscale testing of qualitative hypotheses. *The Annals of Statistics* 29, 124–152.

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

Guerre, E. & P. Lavergne (2002) Minimax rates for nonparametric specification testing in regression models. *Econometric Theory* 18, 1139–1171.

Hall, P. (1991) On convergence rates of suprema. *Probability Theory and Related Fields* 89, 447–455.

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

Ingster, Y. (1987) Asymptotically minimax testing of nonparametric hypotheses. In Y. Prohorov (ed.), *Probability Theory and Mathematical Statistics*. Proceedings of the 4th Vilnuis Conference, VNU Science Press, pp. 553–573.

Ingster, Y. & I. Suslina (2003) *Nonparametric Goodness-of-Fit Testing Under Gaussian Models.* Springer.

Lee, S., K. Song, & Y. Whang (2013) Testing functional inequalities. *Journal of Econometrics* 172, 14–32.

Lepski, O. & V. Spokoiny (1999) Minimax nonparametric hypothesis testing: The case of an inhomogeneous alternative. *Bernoulli* 5, 333–358.

Lepski, O. & A. Tsybakov (2000) Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point. *Probability Theory and Related Fields* 117, 17–48.

Loh, W. (1985) A new method for testing separate families of hypotheses. *Journal of the American Statistical Association* 80, 362–368.

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

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

Ponomareva, M. (2010). Inference in models defined by conditional moment inequalities with continuous covariates. Unpublished manuscript.

Rice, J. (1984). Bandwidth choice for nonparametric Kernel regression. *The Annals of Statistics* 12, 1215–1230.

Romano, J., A. Shaikh, & M. Wolf (2014) A practical two-step method for testing moment inequalities. *Econometrica* 82, 1979–2002.

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

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

Appendix A: Implementation Details

In this section, I give some practical recommendations on how to choose tuning parameters a, τ_0, c_h, γ, and the kernel K that are required for implementation of the tests developed in this paper (note that γ is only required for implementation of the RMS critical value). First, I suggest setting $a = 1/2$ since this value worked well in my simulations. Second, the choice of τ_0 depends on how many finite moments ε_j's are expected to have conditional on $X_1^n = (X_i)_{i=1}^n$. Indeed, recall that τ_0 have to satisfy $\tau_0 > d/(q-2)$. Therefore, if, for example, A2 is expected to hold with $q = 4$ (four finite moments), one can set $\tau_0 = d/(q-3/2) = 2d/5$. If A2 is expected to hold with $q = 8$ (eight finite moments), one can set $\tau_0 = d/(q-3/2) = 2d/13$. In my Monte Carlo simulation study with $d = 1$, described above, the choice $\tau_0 = 1$ worked well. Third, following A3, I suggest setting $h_{\max} = \max_{1 \leqslant i,l \leqslant n} \|X_i - X_l\|$ and $h_{\min} = c_h^{1/d} h_{\max}(\log n/n)^{1/(2\tau_0+d)}$ with $c_h = 0.2$. Fourth, I suggest setting $\gamma = 0.046 \times n^{-0.17}$. This choice gives $\gamma = 0.018$ when $n = 250$ and $\gamma = 0.016$ when $n = 500$. For these values of n, this choice of γ is consistent with recommendations in some other papers for comparable settings; see, for example, Chernozhukov et al. (2013b). Finally, although many kernels K lead to good results, as a focal point, I suggest setting $K(x) = 1.5\max(1 - 4\|x\|^2, 0)$ for $x \in \mathbb{R}^d$, since this choice of K worked well in my simulations and is also used in some other related papers; see, for example, Lee et al. (2013).

Appendix B: Proofs

This Appendix contains proofs of all results stated in the main text. Section B.1 contains a useful result on design points $(X_i)_{i=1}^n$. Section B.2 gives a proof of consistency of the estimators $\widehat{\Sigma}_i$ described in Section 2.3. I provide the proof because I was not able to find it in the literature in the stated form. Section B.3 derives a bound on the modulus of continuity of the square root operator on the set of symmetric positive semidefinite matrices. Section B.4 explains a useful anti-concentration inequality for the maximum of Gaussian random variables. Section B.5 describes a result on Gaussian random variables that is used in the proof of the lower bound on the minimax rate. Section B.6 develops some preliminary technical results necessary for the proofs of the main theorems. Finally, Section B.7 presents the proofs of the theorems stated in the main text.

In this Appendix, c and C are used as generic strictly positive constants that depend only on c_1, c_h, C_1, a, p, d, q, α, and τ_0 (but not on n). Their values can change from line to line. It is implicitly assumed that c is small and C is large.

B.1. A Result on Design Points $(X_i)_{i=1}^n$

LEMMA B1. *Suppose that A1 holds. Then there exist constants $c, C > 0$ such that for all $\epsilon \in [c_1(\log^2 n/n)^{1/d}, diam(\mathcal{X})]$ and $x \in \mathcal{X}$,*

$$cn\epsilon^d \leqslant \left|\{i = 1, \ldots, n \colon \|X_i - x\| < \epsilon\}\right| \leqslant Cn\epsilon^d$$

with probability at least $1 - Cn^{-c}$.

Proof. Let

$$\mathcal{N} = \left\{\epsilon = 2^{-k}\text{diam}(\mathcal{X}) \colon \epsilon > 8^{-1}c_1(\log^2 n/n)^{1/d},\ k = 0, 1, 2, \ldots\right\}.$$

Also, for all $\epsilon \in \mathcal{N}$, let $N(\epsilon)$ be a minimal ϵ-net covering $\mathcal{X}$, that is, $N(\epsilon) = \{x_1, \ldots, x_k\} \subset \mathcal{X}$ such that (i) for all $x \in \mathcal{X}$, there is $x' \in N(\epsilon)$ satisfying $\|x - x'\| < \epsilon$, and (ii) there is no set with property (i) of cardinality strictly smaller than k. Since $\text{diam}(\mathcal{X}) \leqslant C_1$ by A1, it follows that $|N(\epsilon)| \leqslant C/\epsilon^d$ for all $\epsilon \in \mathcal{N}$. Thus,

$$\sum_{\epsilon \in \mathcal{N}} |N(\epsilon)| \leqslant n^C. \tag{B.1}$$

Further, fix $\epsilon \in \mathcal{N}$ and $x \in N(\epsilon)$. For $i = 1, \ldots, n$, define $Z_i = 1$ if $\|X_i - x\| < 2\epsilon$ and 0 otherwise. Then $|Z_i - \mathrm{E}[Z_i]| \leqslant 1$ and

$$\mathrm{E}\left[|Z_i - \mathrm{E}[Z_i]|^2\right] \leqslant \mathrm{E}[|Z_i|^2] = \mathrm{E}[Z_i] = \mathrm{P}(\|X_i - x\| < 2\epsilon) \leqslant C\epsilon^d \tag{B.2}$$

since the pdf of X is bounded from above by A1. Therefore, by Bernstein's inequality (see Corollary 2.11 in Boucheron, Lugosi, and Massart, 2013), for any $t > 0$,

$$\mathrm{P}\left(\left|\sum_{i=1}^{n}(Z_i - \mathrm{E}[Z_i])\right| > t\right) \leqslant 2\exp\left(-\frac{t^2}{Cn\epsilon^d + Ct}\right). \tag{B.3}$$

Also,

$$cn\epsilon^d \leqslant \sum_{i=1}^{n}\mathrm{E}[Z_i] \leqslant Cn\epsilon^d. \tag{B.4}$$

Indeed, the upper bound in this inequality follows from (B.2), and the lower bound holds because

$$\mathrm{E}[Z_i] = \mathrm{P}\big(\|X_i - x\| < 2\epsilon\big) \geqslant \mathrm{P}\big(\|X_i - x\| < 1 \wedge (2\epsilon)\big) \geqslant c \cdot (1 \wedge (2\epsilon)) \geqslant c\epsilon,$$

where the second inequality follows from the conditions that

$$\lambda\big(B(x, 1 \wedge (2\epsilon)) \cap \mathcal{X}\big) \geqslant c_1 \lambda\big(B(x, 1 \wedge (2\epsilon))\big)$$

and that the pdf of X is bounded below from zero, which are imposed in A1, and the third inequality follows from $\epsilon \leqslant \mathrm{diam}(\mathcal{X}) \leqslant C_1$.

Now, applying (B.3) with $t = 2^{-1}\sum_{i=1}^{n}\mathrm{E}[Z_i]$ and using (B.4) gives

$$\mathrm{P}\left(cn\epsilon^d \leqslant \sum_{i=1}^{n} Z_i \leqslant Cn\epsilon^d\right) \geqslant 1 - 2\exp(-cn\epsilon^d) \geqslant 1 - 2\exp(-c\log^2 n).$$

Combining this inequality with (B.1) and using the union bound shows that with probability at least $1 - Cn^{-c}$,

$$cn\epsilon^d \leqslant \big|\{i = 1, \ldots, n \colon \|X_i - x\| < 2\epsilon\}\big| \leqslant Cn\epsilon^d$$

for all $\epsilon \in \mathcal{N}$ and $x \in N(\epsilon)$. The asserted claim now follows by noting that for all $\epsilon \in [c_1(\log^2 n/n)^{1/d}, \mathrm{diam}(\mathcal{X})]$ and $x \in \mathcal{X}$, there exist $\epsilon' \in \mathcal{N}$ with $\epsilon' \in (\epsilon/8, \epsilon/4]$, $\epsilon'' \in \mathcal{N}$ with $\epsilon'' \in [\epsilon, 2\epsilon)$, $x' \in N(\epsilon')$, and $x'' \in N(\epsilon'')$ such that

$$B(x', 2\epsilon') \subset B(x, \epsilon) \subset B(x'', 2\epsilon'').$$

This completes the proof of the lemma. ■

B.2. A Result on Estimation of Σ_i's

LEMMA B2. *For $i = 1, \ldots, n$, let $\widehat{\Sigma}_i$ be an estimator of Σ_i described in Section 2.3. Suppose that A1, A2, and A7 hold. In addition, assume that (i) $\|\Sigma(x') - \Sigma(x'')\| \leqslant C_1\|x' - x''\|$ for all $x', x'' \in \mathcal{X}$, and (ii) $b = b_n$ is such that $b \leqslant C_1 n^{-c_1}$ and $n^{(q-4)/(q+4)}b^d \geqslant c_1 n^{c_1}$. Then A5 holds with c_1 and C_1 replaced by c_Σ and C_Σ, respectively, where c_Σ and C_Σ depend only on the constants appearing in A1, A2, and A7.*

Comment B1. In some applications, the functions f_j may admit discontinuities and violate A7, so I want to emphasize that the lemma can be extended to cover some simple discontinuities. For example, assume that the support $\mathcal{X}$ of the random vector X can be partitioned into a finite number of subsets, $\mathcal{X} = \cup_{g=1}^{G} \mathcal{X}_g$, such that A7 holds on each of these subsets in the sense that for all $g = 1, \ldots, G$, $|f_j(x') - f_j(x'')| \leqslant L\|x' - x''\|^{\tau}$ for all $j = 1, \ldots, p$ and $x', x'' \in \mathcal{X}_g$. Also, assume that the estimators $\widehat{\Sigma}_i$ are obtained using the algorithm described in Section 2.3 separately for each $\mathcal{X}_g$, that is, for each $i = 1, \ldots, n$ with $X_i \in \mathcal{X}_g$, the estimator $\widehat{\Sigma}_i$ is obtained using only the data (X_l, W_l) for $l = 1, \ldots, n$ such that $X_l \in \mathcal{X}_g$. Finally, assume that for all $g = 1, \ldots, G$, $x \in \mathcal{X}_g$ and $\epsilon \in (0,1)$, $\lambda(B(x,\epsilon) \cap \mathcal{X}_g) \geqslant c_1 \lambda(B(x,\epsilon))$. Then applying the argument in the lemma for each $\mathcal{X}_g$ separately shows that in this case, the asserted claim of the lemma also holds.

Proof. Since all norms on the finite-dimensional linear space are equivalent in the sense that if $\|\cdot\|_1$ and $\|\cdot\|_2$ are two norms on the finite-dimensional linear space $\mathbb{A}$ then $c\|a\|_1 \leqslant \|a\|_2 \leqslant C\|a\|_1$ for some constants $c, C > 0$ uniformly over $a \in \mathbb{A}$, it is enough to prove that

$$\mathrm{P}\left(\max_{i=1,\ldots,n} |\widehat{\Sigma}_{i,j_1 j_2} - \Sigma_{i,j_1 j_2}| > C_{\Sigma} n^{-c_{\Sigma}}\right) \leqslant C_{\Sigma} n^{-c_{\Sigma}}$$

for all $j_1, j_2 = 1, \ldots, p$, sufficiently small constant c_{Σ}, and sufficiently large constant C_{Σ}. The proof will be given for $j_1 = j_2 = 1$. The result for all other j_1's and j_2's follows from the same argument. To simplify notation, I 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 $\varepsilon_{i,1}$, respectively.

Let $\delta = q - 4 > 0$ and $M = M_n = n^{1/(4+\delta/2)}$. Also, let $X_1^n = (X_i)_{i=1}^n$. For $i = 1, \ldots, n$, denote $\tilde{\varepsilon}_i = \varepsilon_i I\{|\varepsilon_i| \leqslant M\}$. Since $\mathrm{E}[|\varepsilon_i|^{4+\delta} \mid X_1^n] \leqslant C_1^{4+\delta}$ by A2, it follows that

$$\begin{aligned}\mathrm{E}\Big[\max_{i=1,\ldots,n} |\varepsilon_i| \mid X_1^n\Big] &\leqslant \Big(\mathrm{E}\Big[\max_{i=1,\ldots,n} |\varepsilon_i|^{4+\delta} \mid X_1^n\Big]\Big)^{1/(4+\delta)} \\ &\leqslant \Big(\mathrm{E}\Big[\textstyle\sum_{i=1}^n |\varepsilon_i|^{4+\delta} \mid X_1^n\Big]\Big)^{1/(4+\delta)} \leqslant C_1 n^{1/(4+\delta)}.\end{aligned}$$

Then Markov's inequality gives

$$\mathrm{P}\left(\max_{i=1,\ldots,n} |\varepsilon_i| > M \mid X_1^n\right) \leqslant \frac{C_1 n^{1/(4+\delta)}}{M} \leqslant C n^{-c}.$$

So,

$$\mathcal{P} = \mathrm{P}\left(\max_{i=1,\ldots,n} |\tilde{\varepsilon}_i - \varepsilon_i| > 0 \mid X_1^n\right) \leqslant C n^{-c}.$$

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

$$\bar{\Sigma}_i = \frac{1}{2|J(i)|} \sum_{j \in J(i)} \big(\tilde{Y}_{\mathcal{K}(j)} - \tilde{Y}_j\big)\big(\tilde{Y}_{\mathcal{K}(j)} - \tilde{Y}_j\big)^T.$$

Then

$$\mathrm{P}\left(\max_{i=1,\ldots,n} |\bar{\Sigma}_i - \widehat{\Sigma}_i| > 0 \mid X_1^n\right) \leqslant \mathcal{P} \leqslant C n^{-c}. \qquad \textbf{(B.5)}$$

Next, for $i = 1, \ldots, n$, denote $\tilde{\Sigma}_i = \mathrm{E}[\tilde{\varepsilon}_i^2 \mid X_1^n]$. Then $\tilde{\Sigma}_i = \Sigma_i - \mathrm{E}[\varepsilon_i^2 I\{|\varepsilon_i| > M\} \mid X_1^n]$. Combining Fubini's theorem and Markov's inequality yields

$$\begin{aligned}\mathrm{E}\Big[\varepsilon_i^2 I\{\varepsilon_i > M\} \mid X_1^n\Big] &= \int_0^\infty \mathrm{P}\Big(\varepsilon_i^2 I\{|\varepsilon_i| > M\} > t \mid X_1^n\Big) dt \\ &\leqslant M^2 \mathrm{P}(|\varepsilon_i| > M \mid X_1^n) + \int_{M^2}^\infty \frac{\mathrm{E}\big[\varepsilon_i^4 \mid X_1^n\big]}{t^2} dt \\ &\leqslant \frac{2\mathrm{E}\big[\varepsilon_i^4 \mid X_1^n\big]}{M^2} \leqslant \frac{2C_1^4}{n^{2/(4+\delta/2)}}.\end{aligned}$$

Thus, denoting $\ell_n = 2C_1^4 / n^{2/(4+\delta/2)} \leqslant Cn^{-c}$, we obtain

$$\max_{i=1,\ldots,n} |\tilde{\Sigma}_i - \Sigma_i| \leqslant \ell_n. \tag{B.6}$$

Further, since $|\Sigma(x') - \Sigma(x'')| \leqslant C_1 \|x' - x''\|$ for all $x', x'' \in \mathcal{X}$ and $b \leqslant C_1 n^{-c_1}$, it follows that

$$\max_{i=1,\ldots,n} \left| \frac{1}{|J(i)|} \sum_{j \in J(i)} \Sigma_j - \Sigma_i \right| \leqslant C_1^2 n^{-c_1},$$

and so (B.6) implies that

$$\max_{i=1,\ldots,n} \left| \frac{1}{|J(i)|} \sum_{j \in J(i)} \tilde{\Sigma}_j - \Sigma_i \right| \leqslant C_1^2 n^{-c_1} + \ell_n. \tag{B.7}$$

Now, combining (B.5) and (B.7) gives

$$\begin{aligned}&\mathrm{P}\Big(\max_{i=1,\ldots,n} |\widehat{\Sigma}_i - \Sigma_i| > C_\Sigma n^{-c_\Sigma} \mid X_1^n\Big) \\ &\quad \leqslant \mathrm{P}\Big(\max_{i=1,\ldots,n} \Big|\bar{\Sigma}_i - \frac{1}{|J(i)|} \textstyle\sum_{j \in J(i)} \tilde{\Sigma}_j\Big| > C_\Sigma n^{-c_\Sigma} - C_1^2 n^{-c_1} - \ell_n \mid X_1^n\Big) + Cn^{-c} \\ &\quad \leqslant \mathrm{P}\Big(\max_{i=1,\ldots,n} \Big|\bar{\Sigma}_i - \frac{1}{|J(i)|} \textstyle\sum_{j \in J(i)} \tilde{\Sigma}_j\Big| > 2^{-1} C_\Sigma n^{-c_\Sigma} \mid X_1^n\Big) + Cn^{-c}\end{aligned}$$

if c_Σ is small enough and C_Σ is large enough. Further, by the union bound,

$$\begin{aligned}&\mathrm{P}\Big(\max_{i=1,\ldots,n} \Big|\bar{\Sigma}_i - \frac{1}{|J(i)|} \textstyle\sum_{j \in J(i)} \tilde{\Sigma}_j\Big| > 2^{-1} C_\Sigma n^{-c_\Sigma} \mid X_1^n\Big) \\ &\quad \leqslant \sum_{i=1}^n \mathrm{P}\Big(\Big|\bar{\Sigma}_i - \frac{1}{|J(i)|} \textstyle\sum_{j \in J(i)} \tilde{\Sigma}_j\Big| > 2^{-1} C_\Sigma n^{-c_\Sigma} \mid X_1^n\Big).\end{aligned}$$

Moreover, for $i = 1, \ldots, n$,

$$\mathrm{P}\left(\Big|\bar{\Sigma}_i - \frac{1}{|J(i)|} \textstyle\sum_{j \in J(i)} \tilde{\Sigma}_j\Big| > 2^{-1} C_\Sigma n^{-c_\Sigma} \mid X_1^n\right) \leqslant P_{i,1} + P_{i,2} + P_{i,3}$$

where

$$P_{i,1} = \mathrm{P}\Big(\frac{1}{2|J(i)|}\textstyle\sum_{j\in J(i)}\big(f(X_{\mathcal{K}(j)}) - f(X_j)\big)^2 > 6^{-1}C_\Sigma n^{-c_\Sigma} \mid X_1^n\Big),$$
$$P_{i,2} = \mathrm{P}\Big(\frac{1}{|J(i)|}\big|\textstyle\sum_{j\in J(i)}\big(f(X_{\mathcal{K}(j)}) - f(X_j)\big)\big(\tilde{\varepsilon}_{\mathcal{K}(j)} - \tilde{\varepsilon}_j\big)\big)\big| > 6^{-1}C_\Sigma n^{-c_\Sigma} \mid X_1^n\Big),$$
$$P_{i,3} = \mathrm{P}\Big(\Big|\frac{1}{2|J(i)|}\textstyle\sum_{j\in J(i)}\big(\tilde{\varepsilon}_{\mathcal{K}(j)} - \tilde{\varepsilon}_j\big)^2 - \frac{1}{|J(i)|}\textstyle\sum_{j\in J(i)}\tilde{\Sigma}_j\Big| > 6^{-1}C_\Sigma n^{-c_\Sigma} \mid X_1^n\Big).$$

Consider $P_{i,1}$. By A7,

$$|f(X_{\mathcal{K}(j)}) - f(X_j)| \leqslant L\|X_{\mathcal{K}(j)} - X_j\|^\tau \leqslant L\|X_{\mathcal{K}(j)} - X_i\|^\tau + L\|X_i - X_j\|^\tau \leqslant 2Lb^\tau.$$

Therefore, since $b \leqslant C_1 n^{-c_1}$, $P_{i,1} = 0$ if c_Σ is small enough and C_Σ is large enough. Next, consider $P_{i,3}$. Note that $P_{i,3} \leqslant P_{i,3,1} + P_{i,3,2}$ where

$$P_{i,3,1} = \mathrm{P}\Big(\frac{1}{|J(i)|}\Big|\textstyle\sum_{j\in J(i)}\big(\tilde{\varepsilon}_j^2 - \tilde{\Sigma}_j\big)\Big| > 12^{-1}C_\Sigma n^{-c_\Sigma} \mid X_1^n\Big),$$
$$P_{i,3,2} = \mathrm{P}\Big(\frac{1}{|J(i)|}\Big|\textstyle\sum_{j\in J(i)}\tilde{\varepsilon}_{\mathcal{K}(j)}\tilde{\varepsilon}_j\Big| > 12^{-1}C_\Sigma n^{-c_\Sigma} \mid X_1^n\Big).$$

Applying Hoeffding's inequality (see Theorem 2.8 in Boucheron et al. (2013)) conditional on X_1^n gives

$$P_{i,3,1} \leqslant 2\exp\big(-72^{-1}C_\Sigma^2 n^{-2c_\Sigma}|J(i)|/M^4\big). \tag{B.8}$$

In addition, since b satisfies $n^{(q-4)/(q+4)}b^d \geqslant c_1 n^{c_1}$, it follows from the definition of $\delta = q-4$ that $n^{\delta/(8+\delta)}b^d \geqslant c_1 n^{c_1}$, and so it follows from Lemma B1 that with probability at least $1 - Cn^{-c}$, $|J(i)| \geqslant cnb^d$ for all $i = 1,\ldots,n$. Thus, it follows from the definition of $M = n^{1/(4+\delta/2)}$ that with probability at least $1 - Cn^{-c}$,

$$|J(i)|/M^4 \geqslant cn^{\delta/(8+\delta)}b^d \geqslant cn^{c_1},$$

and so it follows from (B.8) that with probability at least $1 - Cn^{-c}$,

$$\sum_{i=1}^n P_{i,3,1} \leqslant Cn^{-c}$$

if c_Σ is small enough and C_Σ is large enough. Next, consider $P_{i,3,2}$. Denote $U(i) = \{j \in J(i)\colon\ j < \mathcal{K}(j)\}$. Since $|\tilde{\varepsilon}_j| \leqslant M$ for all $j = 1,\ldots,n$, applying Hoeffding's inequality conditional on $(\tilde{\varepsilon}_j)_{j\in U(i)}$ and X_1^n shows that $\sum_{i=1}^n P_{i,3,2} \leqslant Cn^{-c}$ by the same argument as that used to bound $\sum_{i=1}^n P_{i,3,1}$.

Finally, a similar argument shows that $\sum_{i=1}^n P_{i,2} \leqslant Cn^{-c}$. The asserted claim follows. ■

B.3. Continuity of the Square Root Operator for Matrices

LEMMA B3. *Let A and B be $p \times p$-dimensional symmetric positive semidefinite matrices. Then for the matrix norm $\|\cdot\|$ defined in (3), it follows that $\|A^{1/2} - B^{1/2}\| \leqslant p^{1/2}\|A - B\|^{1/2}$.*

Comment B2. In fact, a better result is proven in Bhatia (1997), Theorem X.1.1, where it is shown that $\|A^{1/2} - B^{1/2}\| \leqslant \|A - B\|^{1/2}$. Since the statement of that theorem is quite abstract and the proof is technically rather involved, I provide a simple proof of the stated lemma here.

Proof. Let $a_1, \ldots, a_p$ and $b_1, \ldots, b_n$ be orthonormal eigenvectors of matrices A and B, respectively. Let $\lambda_1(A), \ldots, \lambda_p(A)$ and $\lambda_1(B), \ldots, \lambda_p(B)$ be corresponding eigenvalues. For $i = 1, \ldots, p$, let $f_{i1}, \ldots, f_{ip}$ be coordinates of a_i in the basis $(b_1, \ldots, b_p)$, so that $\sum_{j=1}^{p} f_{ij}^2 = 1$.

Then for any $i = 1, \ldots, p$,

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

since $\|(A - B)a_i\| \leqslant \|A - B\|\|a_i\| = \|A - B\|$. Also, for $P = A$ or B, $P^{1/2}$ has the same eigenvectors as P with corresponding eigenvalues equal to $\lambda_1^{1/2}(P), \ldots, \lambda_p^{1/2}(P)$. Therefore, for any $i = 1, \ldots, p$,

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

where the second line follows from Jensen's inequality and the inequality derived above.

Now, for any $c \in \mathbb{R}^p$ with $\|c\| = 1$, let $d_1, \ldots, d_p$ be coordinates of c in the basis $(a_1, \ldots, a_p)$. Then

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

since $\sum_{i=1}^{p} d_i^2 = 1$. Thus, $\left\|A^{1/2} - B^{1/2}\right\| \leqslant p^{1/2}\|A - B\|^{1/2}$. This completes the proof of the lemma. ■

B.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 Gaussian random variables derived in Chernozhukov et al. (2014a). Let $(Z_i)_{i=1}^{S}$ be a set of standard Gaussian (possibly correlated) random variables. Define $W = \max_{i=1,\ldots,S} Z_i$, and let $f_W(\cdot)$ denote the pdf of W.

LEMMA B4. $\sup_{w \in \mathbb{R}} f_W(w) \leqslant C \log^{1/2} S$ *for some universal constant* C.

Proof. Theorem 3 in Chernozhukov et al. (2014a) proves that $\sup_{w \in \mathbb{R}} f_W(w) \leqslant C\mathrm{E}[W]$. In addition, since $W = \max_{i=1,\ldots,S} Z_i$ and all Z_i's are standard Gaussian, it follows that $\mathrm{E}[W] \leqslant C \log^{1/2} S$. Combining these bounds gives the asserted claim. ■

B.5. A 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 B5. *Let* $(\xi_n)_{n\geqslant 1}$ *be a sequence of independent standard Gaussian random variables and* $(w_{i,n})_{i=1}^n$, $n \geqslant 1$, *be a triangular array of positive numbers. If* $w_{i,n} \leqslant C\sqrt{\log n}$ *with* $C \in (0,1)$ *for all* $i = 1,\dots,n$ *and* $n \geqslant 1$, *then*

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

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

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

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

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

The asserted claim follows. ■

B.6. Preliminary Technical Results

In this section, I derive some preliminary results that are used in the proofs of the theorems stated in the main text. It is assumed throughout that conditions A1–A6 hold. I will use the following additional notation. Let $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$. Also, let $(\epsilon_i)_{i=1}^n$ be an i.i.d. sequence of standard Gaussian random vectors in $\mathbb{R}^p$ that are independent of the data. For $i = 1,\dots,n$, denote $\widehat{e}_i = \widehat{\Sigma}_i^{1/2}\epsilon_i$ and $e_i = \Sigma_i^{1/2}\epsilon_i$. Note that $\widehat{e}_i$ is equal in distribution to $\widetilde{Y}_i \sim N(0_p, \widehat{\Sigma}_i)$. Moreover, let

$$\varepsilon_{i,j,h} = \sum_{l=1}^n w_h(X_i,X_l)\varepsilon_{l,j}, \qquad f_{i,j,h} = \sum_{l=1}^n w_h(X_i,X_l)f_j(X_l),$$

$$e_{i,j,h} = \sum_{l=1}^n w_h(X_i,X_l)e_{l,j}, \qquad \widehat{e}_{i,j,h} = \sum_{l=1}^n w_h(X_i,X_l)\widehat{e}_{l,j},$$

$$T^{PIA} = \max_{(i,j,h)\in\mathcal{S}} \frac{\widehat{e}_{i,j,h}}{\widehat{V}_{j,h}(X_i)}, \qquad T^{PIA,0} = \max_{(i,j,h)\in\mathcal{S}} \frac{e_{i,j,h}}{V_{j,h}(X_i)}.$$

Note that T^{PIA} is equal in distribution to the simulated statistic for the plugin critical value. Finally, for $\beta \in (0,1)$, let $c_{1-\beta}^{PIA,0}$ denote the $(1-\beta)$ quantile of the conditional distribution of $T^{PIA,0}$ given $X_1^n = (X_i)_{i=1}^n$.

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

LEMMA B6. *There exist constants $c, C > 0$ such that with probability at least $1 - Cn^{-c}$,*

$$w_h(X_i, X_l) \leqslant C/(nh^d), \quad \text{for all } i, l = 1, \ldots, n, \text{ and } h \in \mathcal{H},$$

and

$$c/\sqrt{nh^d} \leqslant V_{j,h}(X_i) \leqslant C/\sqrt{nh^d}, \quad \text{for all } (i, j, h) \in \mathcal{S}.$$

Proof. For $h > 0$ and $x \in \mathcal{X}$, let $M_h(x) = |\{i = 1, \ldots, n \colon \|X_i - x\| < h\}|$. By A3, $h_{\max}$ and $h_{\min}$ satisfy $h_{\max} \leqslant \operatorname{diam}(\mathcal{X})$ and with probability at least $1 - Cn^{-c}$, $h_{\min} \geqslant 2c_1(\log^2 n/n)^{1/d}$. Therefore, by Lemma B1, with probability at least $1 - Cn^{-c}$,

$$cnh^d \leqslant M_{h/2}(X_i) \leqslant M_h(X_i) \leqslant Cnh^d$$

for all $i = 1, \ldots, n$ and $h \in \mathcal{H}$. Hence, by A4, with probability at least $1 - Cn^{-c}$,

$$cnh^d \leqslant c_1 M_{h/2}(X_i) \leqslant \sum_{l=1}^{n} K\left(\frac{X_i - X_l}{h}\right) \leqslant C_1 M_h(X_i) \leqslant Cnh^d$$

and

$$cnh^d \leqslant \sum_{l=1}^{n} K^2\left(\frac{X_i - X_l}{h}\right) \leqslant Cnh^d$$

for all $i = 1, \ldots, n$ and $h \in \mathcal{H}$. In addition, $K((X_i - X_l)/h) \leqslant C_1$ for all $l = 1, \ldots, n$, and so with probability at least $1 - Cn^{-c}$,

$$w_h(X_i, X_l) = \frac{K_h(X_i - X_l)}{\sum_{k=1}^{n} K_h(X_i - X_k)} = \frac{K((X_i - X_l)/h)}{\sum_{k=1}^{n} K((X_i - X_k)/h)} \leqslant \frac{C}{nh^d}$$

for all $i, l = 1, \ldots, n$ and $h \in \mathcal{H}$, which gives the first asserted claim.

To prove the second asserted claim, note that by A2, since $\sum_{l=1}^{n} w_h(X_i, X_l) = 1$, with probability at least $1 - Cn^{-c}$,

$$\begin{aligned} V_{j,h}(X_i) &= \left(\textstyle\sum_{l=1}^{n} w_h^2(X_i, X_l)\Sigma_{l,jj}\right)^{1/2} \leqslant C_1\left(\textstyle\sum_{l=1}^{n} w_h^2(X_i, X_l)\right)^{1/2} \\ &\leqslant C_1 \max_{l=1,\ldots,n} w_h^{1/2}(X_i, X_l) \leqslant \frac{C}{\sqrt{nh^d}} \end{aligned}$$

and

$$V_{j,h}(X_i) \geqslant c_1\left(\sum_{l=1}^{n} w_h^2(X_i, X_l)\right)^{1/2} \geqslant \frac{c}{nh^d}\left(\sum_{l=1}^{n} K^2\left(\frac{X_i - X_l}{h}\right)\right)^{1/2} \geqslant \frac{c}{\sqrt{nh^d}}$$

for all $(i, j, h) \in \mathcal{S}$. This completes the proof of the lemma. ■

LEMMA B7. *There exist constants $c, C > 0$ such that*

$$\mathrm{E}\Big[\max_{(i,j,h)\in\mathcal{S}} |e_{i,j,h}/V_{j,h}(X_i)| \mid \mathcal{D}_n\Big] \leqslant C\log^{1/2} n, \tag{B.9}$$

$$\mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}} |e_{i,j,h}/V_{j,h}(X_i)| > C\log^{1/2} n \mid \mathcal{D}_n\Big) \leqslant Cn^{-c}, \tag{B.10}$$

$$c^{PIA,0}_{1-\beta} \leqslant C\log^{1/2} n/\beta, \quad \textit{for all } \beta\in(0,1). \tag{B.11}$$

Proof. The first asserted claim holds since for all $(i,j,h)\in\mathcal{S}$, the random variables $e_{i,j,h}/V_{j,h}(X_i)$ are standard Gaussian conditional on $\mathcal{D}_n$ and $\log|\mathcal{S}| \leqslant C\log n$. The second asserted claim follows from the first one and Borell's inequality (see, for example, Proposition A.2.1 in Van der Vaart and Wellner, 1996). The third asserted claim follows from the first one and Markov's inequality. ■

LEMMA B8. *There exist constants $c, C > 0$ such that*

$$\mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}} |\widehat{V}_{j,h}(X_i)/V_{j,h}(X_i) - 1| > Cn^{-c}\Big) \leqslant Cn^{-c},$$

$$\mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}} |V_{j,h}(X_i)/\widehat{V}_{j,h}(X_i) - 1| > Cn^{-c}\Big) \leqslant Cn^{-c}.$$

Proof. By A2, for all $(i,j,h)\in\mathcal{S}$,

$$V^2_{j,h}(X_i) = \sum_{l=1}^n w_h^2(X_i,X_l)\Sigma_{l,jj} \geqslant c\sum_{l=1}^n w_h^2(X_i,X_l).$$

In addition,

$$\Big|\widehat{V}^2_{j,h}(X_i) - V^2_{j,h}(X_i)\Big| \leqslant \sum_{l=1}^n w_h^2(X_i,X_l)|\widehat{\Sigma}_{l,jj} - \Sigma_{l,jj}|.$$

So,

$$\begin{aligned}\max_{(i,j,h)\in\mathcal{S}} \Big|\widehat{V}^2_{j,h}(X_i)/V^2_{j,h}(X_i) - 1\Big| &\leqslant C\max_{j=1,\dots,p}\max_{l=1,\dots,n}\Big|\widehat{\Sigma}_{l,jj} - \Sigma_{l,jj}\Big| \\ &\leqslant C\max_{l=1,\dots,n}\Big\|\widehat{\Sigma}_l - \Sigma_l\Big\|.\end{aligned}$$

Hence,

$$\mathrm{P}\left(\max_{(i,j,h)\in\mathcal{S}} |\widehat{V}^2_{j,h}(X_i)/V^2_{j,h}(X_i) - 1| > Cn^{-c}\right) \leqslant Cn^{-c}$$

by A5. Combining this result with the inequality $|x-1| \leqslant |x^2-1|$, which holds for all $x > 0$, yields the first asserted claim. The second asserted claim follows from the first one and the inequality $|1/x - 1| < 2|x-1|$, which holds for all $x\in\mathbb{R}$ satisfying $|x-1| < 1/2$. ■

LEMMA B9. *There exist constants $c, C > 0$ and a sequence $(\psi_n)_{n\geqslant 1}$ of positive numbers satisfying $\psi_n \leqslant Cn^{-c}$ such that*

$$\mathrm{P}\Big(c^{PIA,0}_{1-\beta-\psi_n} > c^{PIA}_{1-\beta}\Big) \leqslant Cn^{-c},$$
$$\mathrm{P}\Big(c^{PIA,0}_{1-\beta+\psi_n} < c^{PIA}_{1-\beta}\Big) \leqslant Cn^{-c},$$

for all $\beta \in (0,1)$. (Here, I set $c^{PIA}(\beta) = c^{PIA,0}(\beta) = -\infty$ for $\beta \leqslant 0$ and $c^{PIA}(\beta) = c^{PIA,0}(\beta) = +\infty$ for $\beta \geqslant 1$.)

Proof. Fix $\beta \in (0,1)$. Let $\mathcal{A}$ denote the event that $\|\widehat{\Sigma}_l - \Sigma_l\| \leqslant C_1 n^{-c_1}$ for all $l = 1,\ldots,n$. By A5, the event $\mathcal{A}$ holds with probability at least $1 - Cn^{-c}$. Thus, it is enough to show that $c^{PIA,0}_{1-\beta-\psi_n} \leqslant c^{PIA}_{1-\beta}$ and $c^{PIA,0}_{1-\beta+\psi_n} \geqslant c^{PIA}_{1-\beta}$ on the event $\mathcal{A}$.

For $i = 1,\ldots,n$, let

$$r_i = (r_{i,1},\ldots,r_{i,p})' = \widehat{e}_i - e_i.$$

Also, let

$$p_1 = \max_{(i,j,h)\in\mathcal{S}} \left|\frac{e_{i,j,h}}{V_{j,h}(X_i)}\right| \cdot \max_{(i,j,h)\in\mathcal{S}} \left|\frac{V_{j,h}(X_i)}{\widehat{V}_{j,h}(X_i)} - 1\right|$$

and

$$p_2 = \max_{(i,j,h)\in\mathcal{S}} \left|\frac{\sum_{l=1}^n w_h(X_i,X_l) r_{l,j}}{\widehat{V}_{j,h}(X_i)}\right|.$$

Then, by the triangle inequality,

$$\left|T^{PIA} - T^{PIA,0}\right| \leqslant p_1 + p_2. \tag{B.12}$$

Now, as in the proof of Lemma B8, $\max_{(i,j,h)\in\mathcal{S}} |V_{j,h}(X_i)/\widehat{V}_{j,h}(X_i) - 1| \leqslant Cn^{-c}$ on the event $\mathcal{A}$. Therefore, it follows from (B.10) in Lemma B7 that

$$\mathrm{P}\big(p_1 > Cn^{-c}\log^{1/2} n \mid \mathcal{D}_n\big) \leqslant Cn^{-c} \tag{B.13}$$

on the event $\mathcal{A}$. Further,

$$p_2 \leqslant C \max_{(i,j,h)\in\mathcal{S}} \left|\frac{\sum_{l=1}^n w_h(X_i,X_l) r_{l,j}}{V_{j,h}(X_i)}\right|$$

on the event $\mathcal{A}$. In addition, by Lemma B3, for all $i = 1,\ldots,n$ and $j = 1,\ldots,p$,

$$\begin{aligned}\mathrm{E}\big[r_{i,j}^2 \mid \mathcal{D}_n\big] &\leqslant \mathrm{E}\big[\|r_i\|^2 \mid \mathcal{D}_n\big] \leqslant \mathrm{E}\Big[\big\|\widehat{\Sigma}_i^{1/2} - \Sigma_i^{1/2}\big\|^2 \cdot \|\epsilon_i\|^2 \mid \mathcal{D}_n\Big] \\ &= p\big\|\widehat{\Sigma}_i^{1/2} - \Sigma_i^{1/2}\big\|^2 \leqslant p^2 \|\widehat{\Sigma}_i - \Sigma_i\| \leqslant Cn^{-c}\end{aligned}$$

on the event $\mathcal{A}$. Thus, conditional on $\mathcal{D}_n$, the random variables $\sum_{l=1}^n w_h(X_i,X_l) r_{l,j}/V_{j,h}(X_i)$ are zero-mean Gaussian with variance bounded from above by Cn^{-c} since

$$\max_{(i,j,h)\in\mathcal{S}} \frac{\sum_{l=1}^n w_h^2(X_i,X_l)}{V_{j,h}^2(X_i)} = \max_{(i,j,h)\in\mathcal{S}} \frac{\sum_{l=1}^n w_h^2(X_i,X_l)}{\sum_{l=1}^n w_h^2(X_i,X_l)\Sigma_{l,jj}} \leqslant C$$

by A2. Hence, using the same argument as that in the proof of Lemma B7 gives

$$\mathrm{P}\big(p_2 > Cn^{-c}\log^{1/2} n \mid \mathcal{D}_n\big) \leqslant Cn^{-c} \tag{B.14}$$

on the event $\mathcal{A}$ since $\log|\mathcal{S}| \leqslant C\log n$. Conclude from (B.12), (B.13), and (B.14) that

$$\mathrm{P}\big(|T^{PIA} - T^{PIA,0}| > Cn^{-c}\log^{1/2} n \mid \mathcal{D}_n\big) \leqslant Cn^{-c} \tag{B.15}$$

on the event $\mathcal{A}$.

Next, note that conditional on X_1^n, $T^{PIA,0} = \max_{(i,j,h)\in\mathcal{S}}(e_{i,j,h}/V_{j,h}(X_i))$ is the maximum of $|\mathcal{S}|$ standard Gaussian random variables. Since $\log|\mathcal{S}| \leqslant C\log n$, it follows from Lemma B4 that the conditional density of $T^{PIA,0}$ given X_1^n is bounded from above by $C\log^{1/2} n$, and so for any $\phi \in (0, 1-\beta)$,

$$C\Big(c^{PIA,0}_{1-\beta-\phi/2} - c^{PIA,0}_{1-\beta-\phi}\Big)\log^{1/2} n \geqslant \phi. \tag{B.16}$$

Finally, let $\psi_n = Cn^{-c}$ for some sufficiently small constant c and some sufficiently large constant C. Then on the event $\mathcal{A}$,

$$\begin{aligned}
\mathrm{P}\Big(T^{PIA} \leqslant c^{PIA,0}_{1-\beta-\psi_n} \mid \mathcal{D}_n\Big) &\leqslant \mathrm{P}\Big(T^{PIA,0} \leqslant c^{PIA,0}_{1-\beta-\psi_n} + Cn^{-c}\log^{1/2} n \mid \mathcal{D}_n\Big) + Cn^{-c}\\
&\leqslant \mathrm{P}\Big(T^{PIA,0} \leqslant c^{PIA,0}_{1-\beta-\psi_n/2} \mid \mathcal{D}_n\Big) + Cn^{-c}\\
&= 1-\beta-\psi_n/2 + Cn^{-c}\\
&\leqslant 1-\beta
\end{aligned}$$

by (B.15) and (B.16) since in the definition of ψ_n, c is small enough and C is large enough. This gives the first asserted claim. The second asserted claim follows from a similar argument. This completes the proof of the lemma. ■

LEMMA B10. *There exist constants $c, C > 0$ such that*

$$\Big|\mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/V_{j,h}(X_i)) \leqslant c^{PIA,0}_{1-\beta}\Big) - (1-\beta)\Big| \leqslant Cn^{-c},$$
$$\Big|\mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(-\varepsilon_{i,j,h}/V_{j,h}(X_i)) \leqslant c^{PIA,0}_{1-\beta}\Big) - (1-\beta)\Big| \leqslant Cn^{-c}$$

for all $\beta \in (0,1)$.

Proof. Fix $\beta \in (0,1)$. To prove the asserted claim, I apply Proposition 2.1 in Chernozhukov et al. (2014b). Let $B_n = C/h_{\min}^{d/2}$ for some sufficiently large C. Also, let $\mathcal{A}$ be the event that

$$\frac{w_h(X_i, X_l)\Sigma_{l,jj}^{1/2}}{V_{j,h}(X_i)} \leqslant \frac{C}{\sqrt{nh^d}}$$

for all $(i,j,h) \in \mathcal{S}$, $l = 1,\ldots,n$, and some sufficiently large C. By A2 and Lemma B6, the event $\mathcal{A}$ holds with probability at least $1 - Cn^{-c}$.

Now, for all $(i,j,h) \in \mathcal{S}$,

$$\frac{1}{n}\sum_{l=1}^{n} \mathrm{E}\left[\left|\frac{\sqrt{n}w_h(X_i,X_l)\varepsilon_{l,j}}{V_{j,h}(X_i)}\right|^2 \mid X_1^n\right] = 1.$$

Also, by A2, for all $(i,j,h) \in \mathcal{S}$,

$$\begin{aligned}\frac{1}{n}\sum_{l=1}^{n} \mathrm{E}\left[\left|\frac{\sqrt{n}w_h(X_i,X_l)\varepsilon_{l,j}}{V_{j,h}(X_i)}\right|^3 \mid X_1^n\right] &\leqslant \frac{C}{n}\sum_{l=1}^{n}\left|\frac{\sqrt{n}w_h(X_i,X_l)}{V_{j,h}(X_i)}\right|^3 \\ &\leqslant \frac{C}{n}\sum_{l=1}^{n}\left|\frac{\sqrt{n}w_h(X_i,X_l)\Sigma_{l,jj}^{1/2}}{V_{j,h}(X_i)}\right|^3 \\ &\leqslant C\max_{l=1,\ldots,n}\frac{\sqrt{n}w_h(X_i,X_l)\Sigma_{l,jj}^{1/2}}{V_{j,h}(X_i)} \leqslant \frac{C\sqrt{n}}{\sqrt{nh^d}} \leqslant B_n\end{aligned}$$

on the event $\mathcal{A}$. Similarly, for all $(i,j,h) \in \mathcal{S}$,

$$\frac{1}{n}\sum_{l=1}^{n} \mathrm{E}\left[\left|\frac{\sqrt{n}w_h(X_i,X_l)\varepsilon_{l,j}}{V_{j,h}(X_i)}\right|^4 \mid X_1^n\right] \leqslant B_n^2$$

on the event $\mathcal{A}$. Moreover, by A2, for all $l = 1,\ldots,p$,

$$\begin{aligned}\mathrm{E}\left[\max_{(i,j,h)\in\mathcal{S}}\left|\frac{\sqrt{n}w_h(X_i,X_l)\varepsilon_{l,j}}{V_{j,h}(X_i)}\right|^q \mid X_1^n\right] &\leqslant C\max_{(i,j,h)\in\mathcal{S}}\left|\frac{\sqrt{n}w_h(X_i,X_l)}{V_{j,h}(X_i)}\right|^q \\ &\leqslant C\max_{(i,j,h)\in\mathcal{S}}\left|\frac{\sqrt{n}w_h(X_i,X_l)\Sigma_{l,jj}^{1/2}}{V_{j,h}(X_i)}\right|^q \leqslant B_n^q\end{aligned}$$

on the event $\mathcal{A}$.

Further, note that $h_{\min}$ depends on the data $\mathcal{D}_n$ only via X_1^n. In addition, since $d/(q-2) < \tau_0 \leqslant 1$, it follows from A1 and A3 that with probability at least $1 - Cn^{-c}$,

$$\left(nh_{\min}^d\right)^{-1} \leqslant C_1 n^{-2/q - c_1}. \qquad \textbf{(B.17)}$$

Therefore, applying Proposition 2.1 in Chernozhukov et al. (2014b) conditional on X_1^n shows that

$$\left|\mathrm{P}\left(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/V_{j,h}(X_i)) \leqslant c_{1-\lambda}^{PIA,0} \mid X_1^n\right) - (1-\lambda)\right| \leqslant Cn^{-c}$$

on the intersection of $\mathcal{A}$ and the event (B.17). Since both $\mathcal{A}$ and (B.17) hold with probability at least $1 - Cn^{-c}$, the first asserted claim follows. The second asserted claim follows from a similar argument. This completes the proof of the lemma. ■

LEMMA B11. *There exist constants $c, C > 0$ such that*

$$\mathrm{P}\left(\max_{(i,j,h)\in\mathcal{S}}|\varepsilon_{i,j,h}/V_{j,h}(X_i)| > C\log^{1/2} n\right) \leqslant Cn^{-c},$$

$$\mathrm{P}\left(\max_{(i,j,h)\in\mathcal{S}}|\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i)| > C\log^{1/2} n\right) \leqslant Cn^{-c}.$$

Proof. Set $\beta = \beta_n = Cn^{-c}$ for some sufficiently small c and some sufficiently large C. Then Lemma B10 implies that

$$\mathrm{P}\left(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/V_{j,h}(X_i)) > c^{PIA,0}_{1-\beta}\right) \leqslant Cn^{-c}. \tag{B.18}$$

Further, combining (B.9) in Lemma B7 and Borell's inequality implies that

$$c^{PIA,0}_{1-\beta} \leqslant C\log^{1/2} n \tag{B.19}$$

Combining (B.18) and (B.19) implies the first asserted claim. The second asserted claim follows by noting that

$$\max_{(i,j,h)\in\mathcal{S}}\left|\frac{\varepsilon_{i,j,h}}{\widehat{V}_{j,h}(X_i)}\right| \leqslant \max_{(i,j,h)\in\mathcal{S}}\left|\frac{\varepsilon_{i,j,h}}{V_{j,h}(X_i)}\right| \cdot \max_{(i,j,h)\in\mathcal{S}}\frac{V_{j,h}(X_i)}{\widehat{V}_{j,h}(X_i)}$$

and that

$$\mathrm{P}\left(\max_{(i,j,h)\in\mathcal{S}}\frac{V_{j,h}(X_i)}{\widehat{V}_{j,h}(X_i)} > 1 + Cn^{-c}\right) \leqslant Cn^{-c}$$

by Lemma B8. This completes the proof of the lemma. ■

LEMMA B12. *There exist constants $c, C > 0$ and a sequence $(\psi'_n)_{n\geqslant 1}$ of positive numbers satisfying $\psi'_n \leqslant Cn^{-c}$ such that*

$$\mathrm{P}\left(\max_{(i,j,h)\in\mathcal{S}\setminus\mathcal{S}^D}(\widehat{f}_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)) > 0\right) \leqslant Cn^{-c} \tag{B.20}$$

and

$$\mathrm{P}(\mathcal{S}^D \subset \mathcal{S}^{RMS}) \geqslant 1 - Cn^{-c} \tag{B.21}$$

where

$$\mathcal{S}^D = \mathcal{S}^D_n = \left\{(i,j,h)\in\mathcal{S}\colon\ f_{i,j,h}/V_{j,h}(X_i) > -c^{PIA,0}_{1-\gamma_n-\psi'_n}\right\}.$$

Proof. Recall the sequence ψ_n appearing in the statement of Lemma B9. Also, denote $R = R_n = \max_{(i,j,h)\in\mathcal{S}}(V_{j,h}(X_i)/\widehat{V}_{j,h}(X_i))$. By Lemma B8, $\mathrm{P}(|R-1| > Cn^{-c}) \leqslant Cn^{-c}$. Therefore, there exist sufficiently small c and sufficiently large C such that $\psi'_n = \psi_n \vee (Cn^{-c})$ satisfies

$$\mathrm{P}(|R-1|/\psi'_n > Cn^{-c}) \leqslant Cn^{-c}. \tag{B.22}$$

Now, to prove (B.20), note that by Lemma B10,

$$\left|\mathrm{P}\left(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/V_{j,h}(X_i)) \leqslant c^{PIA,0}_{1-\gamma_n-\psi'_n}\right) - (1-\gamma_n-\psi'_n)\right| \leqslant Cn^{-c}.$$

Therefore, since for any $(i,j,h) \in \mathcal{S}\backslash\mathcal{S}^D$, $f_{i,j,h}/V_{j,h}(X_i) \leqslant -c^{PIA,0}_{1-\gamma_n-\psi'_n}$, it follows that

$$\begin{aligned}
\mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}\backslash\mathcal{S}^D}(\widehat{f}_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)) > 0\Big) &= \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}\backslash\mathcal{S}^D}(\widehat{f}_{j,h}(X_i)/V_{j,h}(X_i)) > 0\Big)\\
&= \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}\backslash\mathcal{S}^D}(f_{i,j,h}/V_{j,h}(X_i) + \varepsilon_{i,j,h}/V_{j,h}(X_i)) > 0\Big)\\
&\leqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}\backslash\mathcal{S}^D}\big(-c^{PIA,0}_{1-\gamma_n-\psi'_n} + \varepsilon_{i,j,h}/V_{j,h}(X_i)\big) > 0\Big)\\
&\leqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/V_{j,h}(X_i)) > c^{PIA,0}_{1-\gamma_n-\psi'_n}\Big)\\
&\leqslant 1-(1-\gamma_n-\psi'_n)+Cn^{-c}\\
&= \gamma_n+\psi'_n+Cn^{-c}.
\end{aligned}$$

Hence, (B.20) follows by noting that $\gamma_n \leqslant C_1 n^{-c_1}$ by A6 and $\psi'_n \leqslant Cn^{-c}$ by construction of ψ'_n.

Next, to prove (B.21), note that by Lemma B9, $\mathrm{P}(c^{PIA,0}_{1-\gamma_n-\psi'_n} > c^{PIA}_{1-\gamma_n}) \leqslant Cn^{-c}$. In addition, for all $x \in (-1,1)$,

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

Therefore,

$$\begin{aligned}
\mathrm{P}\Big(\mathcal{S}^D \subset \mathcal{S}^{RMS}\Big) &= \mathrm{P}\Big(\min_{(i,j,h)\in\mathcal{S}^D}(\widehat{f}_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)) > -2c^{PIA}_{1-\gamma_n}\Big)\\
&\geqslant \mathrm{P}\Big(R\cdot\min_{(i,j,h)\in\mathcal{S}^D}(\widehat{f}_{j,h}(X_i)/V_{j,h}(X_i)) > -2c^{PIA}_{1-\gamma_n}\Big)\\
&\geqslant \mathrm{P}\Big(R\cdot\min_{(i,j,h)\in\mathcal{S}^D}\big(-c^{PIA,0}_{1-\gamma_n-\psi'_n} + \varepsilon_{i,j,h}/V_{j,h}(X_i)\big) > -2c^{PIA}_{1-\gamma_n}\Big)\\
&= \mathrm{P}\Big(\min_{(i,j,h)\in\mathcal{S}^D}(\varepsilon_{i,j,h}/V_{j,h}(X_i)) > c^{PIA,0}_{1-\gamma_n-\psi'_n} - 2c^{PIA}_{1-\gamma_n}/R\Big)\\
&\geqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(-\varepsilon_{i,j,h}/V_{j,h}(X_i)) < -c^{PIA,0}_{1-\gamma_n-\psi'_n} + 2c^{PIA,0}_{1-\gamma_n-\psi'_n}/R\Big) - Cn^{-c}\\
&\geqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(-\varepsilon_{i,j,h}/V_{j,h}(X_i)) < c^{PIA,0}_{1-\gamma_n-\psi'_n}(1-2|R-1|)\Big) - Cn^{-c}.
\end{aligned}$$

Further, by Lemma B7, $c^{PIA,0}_{1-\gamma_n-\psi'_n} \leqslant C\log^{1/2}n/(\gamma_n+\psi'_n) \leqslant C\log^{1/2}n/\psi'_n$. So, by (B.22), with probability at least $1-Cn^{-c}$,

$$c^{PIA,0}_{1-\gamma_n-\psi'_n}(1-2|R-1|) \geqslant c^{PIA,0}_{1-\gamma_n-\psi'_n} - C|R-1|\log^{1/2}n/\psi'_n \geqslant c^{PIA,0}_{1-\gamma_n-\psi'_n} - Cn^{-c}. \tag{B.23}$$

Also, as in the proof of Lemma B9, for any $\beta \in (0, 1-\gamma_n-\psi'_n)$,

$$C\Big(c^{PIA,0}_{1-\gamma_n-\psi'_n} - c^{PIA,0}_{1-\gamma_n-\psi'_n-\beta}\Big)\log^{1/2}n \geqslant \beta.$$

Hence, there exists $\chi_n = Cn^{-c}$ such that the right-hand side of (B.23) is bounded from below by $c^{PIA,0}_{1-\gamma_n-\psi'_n-\chi_n}$. Therefore,

$$\begin{aligned}
\mathrm{P}(\mathcal{S}^D \subset \mathcal{S}^{RMS}) &\geqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(-\varepsilon_{i,j,h}/V_{j,h}(X_i)) \leqslant c^{PIA,0}_{1-\gamma_n-\psi'_n-\chi_n}\Big) - Cn^{-c}\\
&\geqslant 1-\gamma_n-\psi'_n-\chi_n-Cn^{-c}.
\end{aligned}$$

So, (B.21) follows by noting that $\gamma_n + \psi_n' + \chi_n \leqslant Cn^{-c}$. This completes the proof of the lemma. ■

LEMMA B13. *Suppose that $f_j(x) = 0$ for all $x \in \mathcal{X}$ and $j = 1, \dots, p$. Then there exist constants $c, C > 0$ such that*

$$\mathrm{P}(\mathcal{S}^{RMS} = \mathcal{S}) \geqslant 1 - Cn^{-c}.$$

Proof. By Lemma B9, $\mathrm{P}\left(c^{PIA,0}_{1-\gamma_n-\psi_n} > c^{PIA}_{1-\gamma_n}\right) \leqslant Cn^{-c}$. Also, by Lemma B8,

$$\mathrm{P}\left(\max_{(i,j,h)\in\mathcal{S}} (V_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)) \leqslant 1 + Cn^{-c}\right) \geqslant 1 - Cn^{-c}.$$

Further, since $f_j(x) = 0$ for all $x \in \mathcal{X}$ and $j = 1, \dots, p$, it follows that for all $(i, j, h) \in \mathcal{S}$, $\widehat{f}_{j,h}(X_i) = \varepsilon_{i,j,h}$. Moreover, $c^{PIA,0}_{1-\gamma_n-\psi_n} \geqslant 0$ for sufficiently large n. So,

$$\begin{aligned}
\mathrm{P}(\mathcal{S}^{RMS} = \mathcal{S}) &= \mathrm{P}\Big(\min_{(i,j,h)\in\mathcal{S}} (\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i)) > -2c^{PIA}_{1-\gamma_n}\Big) \\
&\geqslant \mathrm{P}\Big(\min_{(i,j,h)\in\mathcal{S}} (\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i)) > -2c^{PIA,0}_{1-\gamma_n-\psi_n}\Big) - Cn^{-c} \\
&\geqslant \mathrm{P}\Big(\min_{(i,j,h)\in\mathcal{S}} (\varepsilon_{i,j,h}/V_{j,h}(X_i)) \max_{(i,j,h)\in\mathcal{S}} (V_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)) > -2c^{PIA,0}_{1-\gamma_n-\psi_n}\Big) - Cn^{-c} \\
&\geqslant \mathrm{P}\Big(\min_{(i,j,h)\in\mathcal{S}} (\varepsilon_{i,j,h}/V_{j,h}(X_i))(1 + Cn^{-c}) > -2c^{PIA,0}_{1-\gamma_n-\psi_n}\Big) - Cn^{-c} \\
&\geqslant \mathrm{P}\Big(\min_{(i,j,h)\in\mathcal{S}} (\varepsilon_{i,j,h}/V_{j,h}(X_i)) > -2c^{PIA,0}_{1-\gamma_n-\psi_n}(1 - Cn^{-c})\Big) - Cn^{-c} \\
&\geqslant \mathrm{P}\Big(\min_{(i,j,h)\in\mathcal{S}} (\varepsilon_{i,j,h}/V_{j,h}(X_i)) > -c^{PIA,0}_{1-\gamma_n-\psi_n}\Big) - Cn^{-c} \\
&= \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}} (-\varepsilon_{i,j,h}/V_{j,h}(X_i)) < c^{PIA,0}_{1-\gamma_n-\psi_n}\Big) - Cn^{-c}.
\end{aligned}$$

Hence, it follows from Lemma B10 that

$$\mathrm{P}\big(\mathcal{S}^{RMS} = \mathcal{S}\big) \geqslant 1 - \gamma_n - \psi_n - Cn^{-c}.$$

The asserted claim follows by noting that $\gamma_n + \psi_n \leqslant Cn^{-c}$. ■

LEMMA B14. *Suppose that for some $\mathcal{J} \subset \{1, \dots, p\}$, it follows that $f_j(x) = 0$ for all $x \in \mathcal{X}$ and $j \in \mathcal{J}$ but $f_j(x) \leqslant -c_1$ for all $x \in \mathcal{X}$ and $j \notin \mathcal{J}$. In addition, suppose, that $\gamma = \gamma_n$ is such that $\log(1/\gamma) \leqslant C_1 \log n$. Then there exist constants $c, C > 0$ such that*

$$\mathrm{P}(\mathcal{S}^{RMS} = \mathcal{S}^{\mathcal{J}}) \geqslant 1 - Cn^{-c}$$

where $\mathcal{S}^{\mathcal{J}} = \{(i, j, h) \in \mathcal{S} \colon j \in \mathcal{J}\}$.

Proof. It follows from Lemma B13 that with probability at least $1 - Cn^{-c}$, $(i, j, h) \in \mathcal{S}^{RMS}$ for all $(i, j, h) \in \mathcal{S}^{\mathcal{J}}$. Thus, it suffices to show that with probability at least $1 - Cn^{-c}$, $(i, j, h) \notin \mathcal{S}^{RMS}$ for all $(i, j, h) \notin \mathcal{S}^{\mathcal{J}}$. To this end, note that

$$\mathrm{P}\Big(\max_{(i,j,h)\notin\mathcal{S}^{\mathcal{J}}} \big(\widehat{f}_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)\big) \leqslant -2c^{PIA}_{1-\gamma_n}\Big)$$

$$= \mathrm{P}\Big(\max_{(i,j,h)\notin\mathcal{S}^{\mathcal{J}}}\big(f_{i,j,h}/\widehat{V}_{j,h}(X_i)+\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i)\big)\leqslant -2c^{PIA}_{1-\gamma_n}\Big)$$
$$\geqslant \mathrm{P}\Big(\max_{(i,j,h)\notin\mathcal{S}^{\mathcal{J}}}\big(-c\sqrt{nh^d}+\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i)\big)\leqslant -2c^{PIA}_{1-\gamma_n}\Big)-Cn^{-c}$$
$$\geqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}\big(\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i)\big)\leqslant c\sqrt{nh^d_{\min}}-2c^{PIA}_{1-\gamma_n}\Big)-Cn^{-c}\geqslant 1-Cn^{-c},$$

where the third line follows from Lemmas B6 and B8 and the fourth line from Lemma B11, the fact that with probability at least $1-Cn^{-c}$, $nh^d_{\min}\geqslant cn^c$, and the observation that $c^{PIA}_{1-\gamma_n}\leqslant C\log^{1/2}n$, which holds by Borell's inequality applied conditional on X_1^n since $\log(1/\gamma)=\log(1/\gamma_n)\leqslant C_1\log n$. This completes the proof of the lemma. ■

LEMMA B15. *There exist constants* $c,C>0$ *such that*

$$\mathrm{P}\Big(c^{PIA}_{1-\alpha}>C\log^{1/2}n\Big)\leqslant Cn^{-c},$$
$$\mathrm{P}\Big(c^{RMS}_{1-\alpha}>C\log^{1/2}n\Big)\leqslant Cn^{-c}.$$

Proof. Since $\mathcal{S}^{RMS}\subset\mathcal{S}$, 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 sufficiently large n, and so Lemma B9 implies that $\mathrm{P}(c^{PIA,0}_{1-\alpha/2}<c^{PIA}_{1-\alpha})\leqslant Cn^{-c}$. In addition, $c^{PIA,0}_{1-\alpha/2}\leqslant C\log^{1/2}n$ by (B.11) in Lemma B7. Combining these bounds yields the first asserted claim and completes the proof of the lemma. ■

LEMMA B16. *Recall the set* $\mathcal{S}^D$ *appearing in Lemma B12. For* $\beta\in(0,1)$, *define* $c^D_{1-\beta}$ *as the* $(1-\beta)$ *quantile of the conditional distribution of* $\max_{(i,j,h)\in\mathcal{S}^D}(\widehat{e}_{i,j,h}/\widehat{V}_{j,h}(X_i))$ *given* $\mathcal{D}_n$. *Also, define* $c^{D,0}_{1-\beta}$ *as the* $(1-\beta)$ *quantile of the conditional distribution of* $\max_{(i,j,h)\in\mathcal{S}^D}(e_{i,j,h}/V_{j,h}(X_i))$ *given* X_1^n. *(If the set* $\mathcal{S}^D$ *is empty, define* $c^D_{1-\beta}=c^{D,0}_{1-\beta}=\infty$.*) Then there exist constants* $c,C>0$ *and a sequence* $(\varphi_n)_{n\geqslant1}$ *of positive numbers satisfying* $\varphi_n\leqslant Cn^{-c}$ *such that*

$$\mathrm{P}\Big(c^{D,0}_{1-\beta-\varphi_n}>c^D_{1-\beta}\Big)\leqslant Cn^{-c},$$
$$\mathrm{P}\Big(c^{D,0}_{1-\beta+\varphi_n}<c^D_{1-\beta}\Big)\leqslant Cn^{-c},$$

for all $\beta\in(0,1)$.

Proof. The proof is the same as that given for Lemma B9. ■

LEMMA B17. *There exist constants* $c,C>0$ *such that*

$$\left|\mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}^D}(\varepsilon_{i,j,h}/V_{j,h}(X_i))\leqslant c^{D,0}_{1-\beta}\Big)-(1-\beta)\right|\leqslant Cn^{-c},$$
$$\left|\mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}^D}(-\varepsilon_{i,j,h}/V_{j,h}(X_i))\leqslant c^{D,0}_{1-\beta}\Big)-(1-\beta)\right|\leqslant Cn^{-c}$$

for all $\beta\in(0,1)$.

Proof. The proof is the same as that given for Lemma B10. ■

B.7. Proofs of Theorems

Proof of Theorem 1. Assume that H_0 holds. First, consider the plugin critical value $c^{PIA}_{1-\alpha}$. Since the kernel K is positive by A4, it follows that for all $(i,j,h) \in \mathcal{S}$, $f_{i,j,h} \leqslant 0$. Also, by Lemma B9, $\mathrm{P}\big(c^{PIA,0}_{1-\alpha-\psi_n} > c^{PIA}_{1-\alpha}\big) \leqslant Cn^{-c}$. In addition, $c^{PIA,0}_{1-\alpha-\psi_n} \geqslant 0$ for sufficiently large n since $\alpha < 1/2$. Moreover, by Lemma B8,

$$\mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(V_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)) \leqslant 1 + Cn^{-c}\Big) \geqslant 1 - Cn^{-c}.$$

So,

$$\begin{aligned}\mathrm{P}\Big(T \leqslant c^{PIA}_{1-\alpha}\Big) &= \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}\big(\widehat{f}_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)\big) \leqslant c^{PIA}_{1-\alpha}\Big)\\ &\geqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i)) \leqslant c^{PIA}_{1-\alpha}\Big)\\ &\geqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i)) \leqslant c^{PIA,0}_{1-\alpha-\psi_n}\Big) - Cn^{-c}\\ &\geqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/V_{j,h}(X_i))\max_{(i,j,h)\in\mathcal{S}}(V_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)) \leqslant c^{PIA,0}_{1-\alpha-\psi_n}\Big) - Cn^{-c}\\ &\geqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/V_{j,h}(X_i))(1+Cn^{-c}) \leqslant c^{PIA,0}_{1-\alpha-\psi_n}\Big) - Cn^{-c}\\ &\geqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/V_{j,h}(X_i)) + Cn^{-c} \leqslant c^{PIA,0}_{1-\alpha-\psi_n}\Big) - Cn^{-c}, \qquad \textbf{(B.24)}\end{aligned}$$

where the last line follows from $\mathrm{P}(\max_{(i,j,h)\in\mathcal{S}}|\varepsilon_{i,j,h}/V_{j,h}(X_i)| > C\log^{1/2}n) \leqslant Cn^{-c}$, which is established in Lemma B11.

Next, it follows as in the proof of Lemma B9 that there exists $\chi_n = Cn^{-c}$ such that the expression in (B.24) is bounded from below by

$$\mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/V_{j,h}(X_i)) \leqslant c^{PIA,0}_{1-\alpha-\psi_n-\chi_n}\Big) - Cn^{-c},$$

which is in turn bounded from below by $1-\alpha-\psi_n-\chi_n-Cn^{-c} \geqslant 1-\alpha-Cn^{-c}$ by Lemma B10 and since $\psi_n+\chi_n \leqslant Cn^{-c}$. Hence, (4) with $P = PIA$ follows.

Second, consider the RMS critical value $c^{RMS}_{1-\alpha}$. Recall the set $\mathcal{S}^D$ appearing in Lemma B12 and the random variable $c^D_{1-\alpha}$ appearing in Lemma B16. Then it follows from (B.21) in Lemma B12 that $\mathrm{P}(c^D_{1-\alpha} > c^{RMS}_{1-\alpha}) \leqslant Cn^{-c}$. In addition, by (B.20) in Lemma B12, $\mathrm{P}(\max_{(i,j,h)\in\mathcal{S}\setminus\mathcal{S}^D}\widehat{f}_{j,h}(X_i)/\widehat{V}_{j,h}(X_i) > 0) \leqslant Cn^{-c}$. Moreover, $c^D_{1-\alpha} > 0$ since $\alpha < 1/2$. So,

$$\begin{aligned}\mathrm{P}\Big(T \leqslant c^{RMS}_{1-\alpha}\Big) &= \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}\big(\widehat{f}_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)\big) \leqslant c^{RMS}_{1-\alpha}\Big)\\ &\geqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}\big(\widehat{f}_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)\big) \leqslant c^{D}_{1-\alpha}\Big) - Cn^{-c}\\ &\geqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}^D}\big(\widehat{f}_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)\big) \leqslant c^{D}_{1-\alpha}\Big) - Cn^{-c}.\end{aligned}$$

From this point, (4) with $P = RMS$ follows from the same argument as that used in the case $P = PIA$ with $\mathcal{S}$ replaced by $\mathcal{S}^D$ and also Lemmas B16 and B17 used instead of Lemmas B9 and B10.

Third, assume that $f_j(x) = 0$ for all $x \in \mathcal{X}$ and $j = 1, \ldots, p$ and consider the plugin critical value $c^{PIA}_{1-\alpha}$. By Lemma B9, $\mathrm{P}\big(c^{PIA,0}_{1-\alpha+\psi_n} < c^{PIA}_{1-\alpha}\big) \leqslant Cn^{-c}$. In addition, by Lemma B8,

$$\mathrm{P}\left(\min_{(i,j,h)\in\mathcal{S}}(V_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)) \geqslant 1 - Cn^{-c}\right) \geqslant 1 - Cn^{-c}.$$

So,

$$\begin{aligned}
\mathrm{P}\Big(T \leqslant c^{PIA}_{1-\alpha}\Big) &= \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}\big(\widehat{f}_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)\big) \leqslant c^{PIA}_{1-\alpha}\Big)\\
&= \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i)) \leqslant c^{PIA}_{1-\alpha}\Big)\\
&\leqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i)) \leqslant c^{PIA,0}_{1-\alpha+\psi_n}\Big) + Cn^{-c}\\
&\leqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/V_{j,h}(X_i)) \min_{(i,j,h)\in\mathcal{S}}(V_{j,h}(X_i)/\widehat{V}_{j,h}(X_i)) \leqslant c^{PIA,0}_{1-\alpha+\psi_n}\Big) + Cn^{-c}\\
&\leqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/V_{j,h}(X_i))(1 - Cn^{-c}) \leqslant c^{PIA,0}_{1-\alpha+\psi_n}\Big) + Cn^{-c}\\
&\leqslant \mathrm{P}\Big(\max_{(i,j,h)\in\mathcal{S}}(\varepsilon_{i,j,h}/V_{j,h}(X_i)) - Cn^{-c} \leqslant c^{PIA,0}_{1-\alpha+\psi_n}\Big) + Cn^{-c}.
\end{aligned}$$

Now, an argument like that used above shows that the last expression is bounded from above by $1 - \alpha + Cn^{-c}$. This gives (5) for $P = PIA$.

Fourth, for the RMS critical value $c^{RMS}_{1-\alpha}$ in the case that $f_j(x) = 0$ for all $x \in \mathcal{X}$ and $j = 1, \ldots, p$, note that by Lemma B13, $\mathrm{P}(\mathcal{S}^{RMS} = \mathcal{S}) \geqslant 1 - Cn^{-c}$ in this case, and so

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

Finally, for the RMS critical value $c^{RMS}_{1-\alpha}$ in the case that $f_j(x) = 0$ for all $x \in \mathcal{X}$ and $j \in \mathcal{J}$ but $f_j(x) \leqslant -c_1$ for all $x \in \mathcal{X}$ and $j \notin \mathcal{J}$, note that by Lemma B14, $\mathrm{P}(\mathcal{S}^{RMS} = \mathcal{S}^{\mathcal{J}}) = 1 - Cn^{-c}$, and so the same argument as that used above with $\mathcal{S}$ replaced by $\mathcal{S}^{\mathcal{J}}$ gives

$$\mathrm{P}\big(T \leqslant c^{RMS}_{1-\alpha}\big) \leqslant 1 - \alpha + Cn^{-c}.$$

This completes the proof of the theorem. ■

Proof of Theorem 2. Let $\rho = \sup_{x\in\mathcal{X}} \max_{j=1,\ldots,p} f_j(x)$. Since $\theta_0 \notin \Theta_I$, it follows that $\rho > 0$, and there exist $x \in \mathcal{X}$ and $j = 1, \ldots, p$ such that $f_j(x) = \mathrm{E}[m_j(X, W, \theta_0) \mid X = x] \geqslant \rho/2$. Since the function f_j is continuous, there exists $h_0 > 0$, which is independent of n, such that $f_j(X_i) \geqslant \rho/4$ for all $i = 1, \ldots, n$ with $\|X_i - x\| < h_0$. Further, with probability at least $1 - Cn^{-c}$, there exists $h \in \mathcal{H}$ such that $ah_0/2 \leqslant h < h_0/2$. Also, by Lemma B1, with probability at least $1 - Cn^{-c}$, on the event that this h exists, there exists $i = 1, \ldots, n$ such that $\|X_i - x\| < h$. For this i, $f_{i,j,h} \geqslant \rho/4$ by A4. Hence, by Lemmas B6 and B8, on the event that these i and h exist, with probability at least $1 - Cn^{-c}$,

$$f_{i,j,h}/\widehat{V}_{j,h}(X_i) \geqslant c\rho\sqrt{nh_0^d},$$

and so

$$\begin{aligned}\mathrm{P}\Big(T \leqslant c^P_{1-\alpha}\Big) &\leqslant \mathrm{P}\Big(f_{i,j,h}/\widehat{V}_{j,h}(X_i)+\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i) \leqslant c^P_{1-\alpha}\Big)+Cn^{-c}\\ &\leqslant \mathrm{P}\Big(c\rho\sqrt{nh_0^d}+\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i) \leqslant c^P_{1-\alpha}\Big)+Cn^{-c}\\ &= \mathrm{P}\Big(-\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i) \geqslant c\rho\sqrt{nh_0^d}-c^P_{1-\alpha}\Big)+Cn^{-c} \leqslant Cn^{-c}\end{aligned}$$

where the last inequality follows from Lemmas B11 and B15. This completes the proof of the theorem. ■

Proof of Theorem 3. To prove this theorem, I construct $x \in \mathcal{X}$, $j = 1,\ldots,p$, $h_0 > 0$, $h \in \mathcal{H}$, and $i = 1,\ldots,n$ applying the argument in the proof of Theorem 2 with the functions f_j replaced by the functions f_j^0. Then it follows that

$$\begin{aligned}\mathrm{P}\Big(T \leqslant c^P_{1-\alpha}\Big) &\leqslant \mathrm{P}\Big(f_{i,j,h}/\widehat{V}_{j,h}(X_i)+\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i) \leqslant c^P_{1-\alpha}\Big)+Cn^{-c}\\ &\leqslant \mathrm{P}\Big(ca_n\rho_0\sqrt{nh_0^d}+\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i) \leqslant c^P_{1-\alpha}\Big)+Cn^{-c}\\ &= \mathrm{P}\Big(-\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i) \geqslant ca_n\rho_0\sqrt{nh_0^d}-c^P_{1-\alpha}\Big)+Cn^{-c} \leqslant Cn^{-c}.\end{aligned}$$

Now, since $a_n(n/\log n)^{1/2} \to \infty$, it follows that $a_n\rho_0(nh_0^d)^{1/2} = \beta_n \log^{1/2} n$ for some sequence of positive numbers $(\beta_n)_{n\geqslant 1}$ satisfying $\beta_n \to \infty$. Hence, the asserted claim follows from Lemmas B11 and B15. ■

Proof of Theorem 4. Suppose that (9) does not hold. Then there exist $\beta > 0$ and a sequence $(\theta_{0,n})_{n\geqslant 1}$ such that $\theta_{0,n} \in \Theta_{a_n}$ for all $n \geqslant 1$ but

$$\mathrm{P}\Big(T(\theta_{0,n}) > c^P_{1-\alpha}(\theta_{0,n})\Big) \leqslant 1-\beta, \quad \text{for infinitely many } n.$$

I will show that this is not possible. In particular, fix any sequence $(\theta_{0,n})_{n\geqslant 1}$ such that $\theta_{0,n} \in \Theta_{a_n}$ for all $n \geqslant 1$. I will show that

$$\mathrm{P}\Big(T(\theta_{0,n}) > c^P_{1-\alpha}(\theta_{0,n})\Big) \to 1.$$

To this end, observe that for all $n \geqslant 1$, there exist $x = x_n \in \mathcal{X}$ and $j = j_n = 1,\ldots,p$ such that

$$f_{j,n}(x) = \mathrm{E}[m_j(X,W,\theta_{0,n}) \mid X = x] \geqslant a_n/2.$$

Further, let $h_{0,n} = (\log n/n)^{1/(2\tau+d)}$. By A7, $f_{j,n} \in \mathcal{F}(\tau, L)$, and so for all $i = 1,\ldots,n$ with $\|X_i - x\| < h_{0,n}$, it follows that

$$f_{j,n}(X_i) \geqslant f_{j,n}(x) - Lh_{0,n}^\tau \geqslant a_n/2 - Lh_{0,n}^\tau \geqslant a_n/4$$

for all sufficiently large n since $a_n(n/\log n)^{\tau/(2\tau+d)} \to \infty$. Next, since $\tau > \tau_0$, it follows that with probability at least $1 - Cn^{-c}$, there exists $h = h_n \in \mathcal{H} = \mathcal{H}_n$ such that $ah_{0,n}/2 \leqslant h < h_{0,n}/2$. As in the proof of Theorem 2, on the event that this h exists, with probability at least $1 - Cn^{-c}$, there exists $i = 1,\ldots,n$ such that

$f_{i,j,h}/\widehat{V}_{j,h} \geqslant c a_n \sqrt{n h_{0,n}^d}$.

Hence, using the same argument as that in the proof of Theorem 2 gives

$$\mathrm{P}\Big(T(\theta_{0,n} \leqslant c_{1-\alpha}^{P}(\theta_{0,n})\Big) \leqslant \mathrm{P}\Big(-\varepsilon_{i,j,h}/\widehat{V}_{j,h}(X_i) \geqslant c a_n \sqrt{n h_{0,n}^d} - c_{1-\alpha}^{P}(\theta_{0,n})\Big) + C n^{-c} \leqslant C n^{-c}$$

where the second inequality follows from Lemmas B11 and B15 since

$$a_n \sqrt{n h_{0,n}^d} = a_n (n/\log n)^{\tau/(2\tau+d)} \log^{1/2} n$$

and $a_n (n/\log n)^{\tau/(2\tau+d)} \to \infty$. This completes the proof of the theorem. ■

Proof of Theorem 5. For $h > 0$, let $q_h : \mathbb{R}^d \to \mathbb{R}$ be a function given by $q_h(x) = L|h|^\tau - L\|x\|^\tau$ if $\|x\| \leqslant h$ and $q_h(x) = 0$ otherwise. Recall that $a_n(n/\log n)^{\tau/(2\tau+d)} \to 0$, so that there exists a sequence of positive numbers $(\beta_n)_{n\geqslant 1}$ such that $a_n = \beta_n^\tau (\log n/n)^{\tau/(2\tau+d)}$ and $\beta_n \to 0$. Set $h_n = \beta_n (\log n/n)^{1/(2\tau+d)}$. Then there exists a set $\{x_1, \ldots, x_{k_n}\} \subset \mathcal{X}$ such that $\|x_{l_1} - x_{l_2}\| > 2h_n$ for all $l_1, l_2 = 1, \ldots, k_n$ with $l_1 \neq l_2$ and $k_n \geqslant (c/h_n)^d$ for some constant c that is independent of n. Further, for $l = 1, \ldots, k_n$, define the vector-valued functions $f^l = f_n^l = (f_{1,n}^l, \ldots, f_{p,n}^l)^T$ as follows. For $j = 2, \ldots, p$ and $x \in \mathcal{X}$, define $f_{j,n}^l(x) = 0$. For $j = 1$ and $x \in \mathcal{X}$, define $f_{j,n}^l(x) = q_{h_n}(x - x_l)$. Also, define $f^0 = (f_1^0, \ldots, f_p^0)^T$ by $f_j^0(x) = 0$ for all $x \in \mathcal{X}$ and $j = 1, \ldots, p$. Note that for all $l = 0, \ldots, k_n$ and $j = 1, \ldots, p$, it follows that $f_j^l \in \mathcal{F}(\tau, L)$. Finally, let $\varepsilon^0 = (\varepsilon_1^0, \ldots, \varepsilon_p^0)^T$ where the random variables ε_j^0 are zero-mean Gaussian with standard deviation c_1, are independent across $j = 1, \ldots, p$, and are independent of X. Note that $\varepsilon^0 \in \mathcal{E}$.

Now, as in the proof of Lemma 6.2 in Dumbgen and Spokoiny (2001), for any sequence of tests $(\phi_n)_{n\geqslant 1}$ such that $\sup_{\varepsilon\in\mathcal{E}} \mathrm{E}_{f^0,\varepsilon}[\phi_n(\mathcal{D}_n)] \leqslant \alpha + o(1)$, it follows that

$$\begin{aligned}\inf_{f\in\times_{j=1}^{p}\mathcal{F}(\tau,L)} \inf_{\varepsilon\in\mathcal{E}} \mathrm{E}_{f,\varepsilon}[\phi_n(\mathcal{D}_n)] - \alpha &\leqslant \min_{l=1,\ldots,k_n} \mathrm{E}_{f^l,\varepsilon^0}[\phi_n(\mathcal{D}_n)] - \mathrm{E}_{f^0,\varepsilon^0}[\phi_n(\mathcal{D}_n)] + o(1)\\ &\leqslant \frac{1}{k_n}\sum_{l=1}^{k_n} \mathrm{E}_{f^l,\varepsilon^0}[\phi_n(D_n)] - \mathrm{E}_{f^0,\varepsilon^0}[\phi_n(\mathcal{D}_n)] + o(1)\end{aligned}$$

Further,

$$\begin{aligned}\frac{1}{k_n}\sum_{l=1}^{k_n} \mathrm{E}_{f^l,\varepsilon^0}[\phi_n(\mathcal{D}_n)] - \mathrm{E}_{f^0,\varepsilon^0}[\phi_n(\mathcal{D}_n)] + o(1) &\leqslant \mathrm{E}_{f^0,\varepsilon^0}\left[\left(\frac{1}{k_n}\sum_{i=1}^{k_n}\frac{dP_l}{dP_0} - 1\right)\phi_n(\mathcal{D}_n)\right] + o(1)\\ &\leqslant \mathrm{E}_{f^0,\varepsilon^0}\left[\left|\frac{1}{k_n}\sum_{i=1}^{k_n}\frac{dP_l}{dP_0} - 1\right|\right] + o(1)\end{aligned}$$

where dP_l/dP_0 denotes the corresponding Radon-Nykodim derivative conditional on $(X_i)_{i=1}^n$. Next, for $l = 1, \ldots, k_n$, let $\omega_l = \omega_{l,n} = c_1(\sum_{i=1}^n (f_1^l(X_i))^2)^{1/2}$ and $\xi_l = \xi_{l,n} = \sum_{i=1}^n f_1^l(X_i)\varepsilon_{i,1}^0/\omega_l$. Then

$$\frac{dP_l}{dP_0} = \exp\left(\omega_l \xi_l - \omega_l^2/2\right).$$

Also, conditional on $(X_i)_{i=1}^n$, the random variables ξ_l are independent standard Gaussian. Further, let $\mathcal{A}$ be the event that for all $\epsilon \in [c_1(\log^2 n/n)^{1/d}, \operatorname{diam}(\mathcal{X})]$ and $x \in \mathcal{X}$,

$$cn\epsilon^d \leqslant \left|\{i = 1, \dots, n \colon \|X_i - x\| < \epsilon\}\right| \leqslant Cn\epsilon^d$$

for sufficiently small c and sufficiently large C. By Lemma B1, the event $\mathcal{A}$ holds with probability $1 - Cn^{-c}$. Now, for all sufficiently large n, on the event $\mathcal{A}$, $w_l \leqslant Cn^{1/2}h_n^{\tau+d/2} < \widetilde{C}\log^{1/2} k_n$ for all $l = 1, \dots, k_n$ and some $\widetilde{C} \in (0,1)$ since $nh_n^{\tau+d/2} = o(\log^{1/2} n)$ and $\log k_n \geqslant c\log n$. Hence, by Lemma B5, on the event $\mathcal{A}$,

$$\mathrm{E}_{f^0,\varepsilon^0}\left[\left|\frac{1}{k_n}\sum_{l=1}^{k_n}\frac{dP_l}{dP_0} - 1\right| \mid (X_i)_{i=1}^n\right] \leqslant Ck_n^{\widetilde{C}^2-1},$$

and outside of this event,

$$\mathrm{E}_{f^0,\varepsilon^0}\left[\left|\frac{1}{k_n}\sum_{l=1}^{k_n}\frac{dP_l}{dP_0} - 1\right| \mid (X_i)_{i=1}^n\right] \leqslant C.$$

Since $k_n \to \infty$ and $\mathrm{P}(\mathcal{A}) \to 0$, conclude that

$$\mathrm{E}_{f^0,\varepsilon^0}\left[\left|\frac{1}{k_n}\sum_{l=1}^{k_n}\frac{dP_l}{dP_0} - 1\right|\right] \to 0.$$

Combining this with the inequalities above gives the asserted claim and completes the proof of the theorem. ■