Econometric Theory, **35**, 2019, 729–776.
doi:10.1017/S0266466618000282

TESTING REGRESSION MONOTONICITY IN ECONOMETRIC MODELS

DENIS CHETVERIKOV
Department of Economics, UCLA

Monotonicity is a key qualitative prediction of a wide array of economic models derived via robust comparative statics. It is therefore important to design effective and practical econometric methods for testing this prediction in empirical analysis. This article develops a general nonparametric framework for testing monotonicity of a regression function. Using this framework, a broad class of new tests is introduced, which gives an empirical researcher a lot of flexibility to incorporate ex ante information she might have. The article also develops new methods for simulating critical values, which are based on the combination of a bootstrap procedure and new selection algorithms. These methods yield tests that have correct asymptotic size and are asymptotically nonconservative. It is also shown how to obtain an adaptive and rate optimal test that has the best attainable rate of uniform consistency against models whose regression function has Lipschitz-continuous first-order derivatives and that automatically adapts to the unknown smoothness of the regression function. Simulations show that the power of the new tests in many cases significantly exceeds that of some prior tests, e.g., that of Ghosal, Sen, and Van der Vaart (2000).

1. INTRODUCTION

The concept of monotonicity plays an important role in economics. For example, in economic theory, monotone comparative statics has been a popular research topic for many years; see Milgrom and Shannon (1994). In industrial organization, lack of monotonicity has been used to detect certain phenomena related to strategic behavior of economic agents that are difficult to detect otherwise; see Ellison and Ellison (2011). In econometric theory, shape restrictions including monotonicity have been argued to be among the most important implications of economic theory that could be used for identification and estimation; see Matzkin (1994). In this article, I develop a general nonparametric framework for testing monotonicity of a regression function.

Date: First version: March 2012. This version: July 9, 2018. Email: chetverikov@econ.ucla.edu. I thank Victor Chernozhukov for encouragement and guidance. I am also grateful to Anna Mikusheva, Isaiah Andrews, Andres Aradillas-Lopez, Moshe Buchinsky, Glenn Ellison, Jin Hahn, Bo Honore, Rosa Matzkin, Jose Montiel Olea, Ulrich Muller, Whitney Newey, Joris Pinkse, and Jack Porter for valuable comments. The first version of the article was presented at the Econometrics lunch at MIT in April, 2012. Address correspondence to Denis Chetverikov, e-mail: chetverikov@econ.ucla.edu.

I consider the model

$$Y = f(X) + \varepsilon, \tag{1}$$

where Y is a scalar dependent random variable, X a scalar covariate, ε an unobserved scalar noise variable satisfying $\mathrm{E}[\varepsilon|X] = 0$ almost surely, and $f(\cdot)$ an unknown function.[1] I am interested in testing the null hypothesis, $\mathcal{H}_0$, that $f(\cdot)$ is nondecreasing against the alternative, $\mathcal{H}_a$, that there are x_1 and x_2 such that $x_1 < x_2$ but $f(x_1) > f(x_2)$. The decision is to be made based on an i.i.d. sample of size n, $\{X_i, Y_i\}_{1 \leqslant i \leqslant n}$ from the distribution of the pair (X, Y). I assume that $f(\cdot)$ is smooth but do not impose any parametric structure on it. I derive a theory that yields tests with the correct asymptotic size. I also show how to obtain consistent tests and how to obtain a test with the optimal rate of uniform consistency against classes of alternatives having continuously differentiable regression functions with Lipschitz-continuous first order derivative. Moreover, the rate optimal test constructed in this article is adaptive in the sense that implementing the test does not require knowing the smoothness of $f(\cdot)$.

Many statistics suitable for testing monotonicity may have highly complicated limit distributions. In some cases, like in the case of the statistic leading to the adaptive and rate optimal test, it is not even clear whether the limit distribution exists. The difficulty here is that the processes underlying the test statistic do not have an asymptotic equicontinuity property, and so classical functional central limit theorems, as presented for example in van der Vaart and Wellner (1996) and Dudley (1999), do not apply.

One of the main contributions of this article is to address these issues by providing bootstrap critical values and proving their validity uniformly over a large class of data generating processes. Several previous articles, for example, Gijbels, Hall, Jones, and Koch (2000), Hall and Heckman (2000), and Ghosal et al. (2000), used specific techniques to prove validity of their tests of monotonicity but it is difficult to generalize their techniques to make them applicable for other tests of monotonicity, in particular for the adaptive and rate optimal test. By contrast, in this article, I introduce a general approach that can be used to prove validity of many different tests of monotonicity. Other shape restrictions, such as concavity and super-modularity, can be tested by procedures similar to those developed in this article.

Another problem is that test statistics studied in this article have some asymptotic distribution when $f(\cdot)$ is constant but may diverge if $f(\cdot)$ is not a constant. This discontinuity implies that for some sequences of models $f(\cdot) = f_n(\cdot)$, the limit distribution depends on the local slope function, which is an unknown infinite-dimensional nuisance parameter that cannot be estimated consistently from the data. A common approach in the literature on testing monotonicity to solve this problem is to calibrate the critical value using the case when the type

[1] The working version of the article, which can be found online, also contains some results for the model with multivariate X's, endogenous X's, and sample selection; see arXiv:1212.6756.

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

I error is maximized (the least favorable model), i.e., the model with constant $f(\cdot)$.[2] By contrast, I develop two selection procedures that estimate the set where $f(\cdot)$ is not strictly increasing, and then adjust the critical value to account for this set. The estimation is conducted so that no violation of the asymptotic size occurs. The critical values obtained using these selection procedures yield important power improvements in comparison with other tests if $f(\cdot)$ is strictly increasing over some subsets of the support of X. The first selection procedure, which is based on the one-step approach, is related to those developed in Chernozhukov, Lee, and Rosen (2013), Andrews and Shi (2010), and Chetverikov (2016), all of which deal with the problem of testing conditional moment inequalities. The second selection procedure is novel and is based on the step-down approach. It is somewhat related to methods developed in Romano and Wolf (2005a) and Romano and Shaikh (2010) but the details are rather different.

Furthermore, an important issue that applies to nonparametric testing in general is how to choose a smoothing parameter for the test. In theory, the optimal smoothing parameter can be derived for many smoothness classes of functions $f(\cdot)$. In practice, however, the smoothness class that $f(\cdot)$ belongs to is usually unknown. I deal with this problem by employing the adaptive testing approach. This approach allows me to obtain tests with good power properties when the information about smoothness of the function $f(\cdot)$ possessed by the researcher is absent or limited. More precisely, I construct a test statistic using many different weighting functions that correspond to many different values of the smoothing parameter so that the distribution of the test statistic is mainly determined by the optimal weighting function. I provide a basic set of weighting functions that yields an adaptive and rate optimal test and show how the researcher can change this set in order to incorporate ex ante information.

The literature on testing monotonicity of a nonparametric regression function is quite large but is not complete. The tests of Gijbels et al. (2000) and Ghosal et al. (2000) are based on the signs of $(Y_{i+k} - Y_i)(X_{i+k} - X_i)$ and may be inconsistent against models with conditional heteroscedasticity; see Section 2 for details. The test of Hall and Heckman (2000) is based on the slopes of local linear estimates of $f(\cdot)$. As explained in Section 2 below, the Hall and Heckman test statistic is contained in the class of the test statistics studied in this article and, in fact, corresponds to the adaptive and rate optimal test with the specific choice of the kernel. Hall and Heckman (2000), however, only established validity of their test for (nonrandom) equidistant X_i's and did not show that their test is adaptive and rate optimal. Moreover, it is not immediately clear how to extend their proof technique to allow for i.i.d. data. My article complements theirs by establishing validity of their test in the i.i.d. setting, improving their critical values, and also establishing adaptivity and rate optimality of their test. Other tests are developed in Schlee (1982), Bowman, Jones, and Gijbels (1998), Dumbgen and Spokoiny

[2] The exception is Wang and Meyer (2011) who use the model with an isotonic estimate of $f(\cdot)$ to simulate the critical value. They do not prove whether their test maintains the required size, however.

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

(2001), Durot (2003), Baraud, Huet, and Laurent (2005), Wang and Meyer (2011), and Gutknecht (2016). The test of Schlee (1982) does not seem to be practical; see Gijbels et al. (2000). The test of Bowman et al. (1998) is known to be inconsistent; see Hall and Heckman (2000). The properties of the test of Durot (2003) are only established for the case of (nonrandom) equidistant X_i's and i.i.d. ε_i's. The test of Baraud et al. (2005) is similar to that of Hall and Heckman (2000) but the validity of the test is only established in the homoscedastic Gaussian noise case.[3] The properties of the test of Wang and Meyer (2011) are not established in the literature. The results on adaptive and rate optimal testing in this article are related to (and inspired by) those in Dumbgen and Spokoiny (2001). An important difference, however, is that Dumbgen and Spokoiny (2001) study the ideal Gaussian white noise model, which allows them to use some fine properties of the Gaussian processes, and do not have to deal with the distributional approximations. Juditsky and Nemirovski (2002) showed in the homoscedastic Gaussian noise case that if the smoothness of the function $f(\cdot)$ is known, then essentially optimal testing can be obtained by considering an optimal estimator of $f(\cdot)$ itself and using a test statistic based on the distance from this estimator to the set of monotone functions.

In a contemporaneous work, Lee, Song, and Whang (2017) derived another approach to testing a general class of functional inequalities, including regression monotonicity, based on L_p-functionals. An advantage of their method is that it can be applied not only to the problem of testing regression monotonicity but also to many other problems, like testing monotonicity of nonparametric quantile functions. A disadvantage of their method, however, is that it yields a nonadaptive test.

Results in this article are also different from those in Romano and Wolf (2013) who also consider the problem of testing monotonicity. In particular, they assume that X is nonstochastic and discrete, which makes their problem semi-parametric and substantially simplifies proving validity of critical values, and they test the null hypothesis that $f(\cdot)$ is *not* weakly increasing against the alternative that it is weakly increasing. Lee, Linton, and Whang (2009) and Delgado and Escanciano (2010) derived tests of stochastic monotonicity, which is a related but different problem. Specifically, stochastic monotonicity means that the conditional cdf of Y given X, $F_{Y|X}(y,x)$, is (weakly) decreasing in x for any fixed y.

I also note that the problem of testing monotonicity is related to but different from the problem of testing conditional moment inequalities, which is concerned with testing the null hypothesis that $f(\cdot)$ is nonnegative against the alternative that there is x such that $f(x) < 0$. Although the latter problem has been extensively studied in the recent econometric literature (see Andrews and Shi, 2010; Chernozhukov et al., 2013; Armstrong, 2014; Armstrong and Chan, 2016; and Chetverikov, 2016 among others), the results from that literature can not be used directly for the former problem. Indeed, under the null hypothesis,

[3] Note that assuming the homoscedastic Gaussian noise eliminates the important problem of finding an appropriate critical value for the test as long as an appropriately studentized test statistic is considered since, in this case, one can simply simulate the critical value from the model with the flat regression function, which gives the least favorable case under the null.

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

the latter problem yields the inequalities $\mathrm{E}[Y_i|X_i] \geqslant 0$, each of which depends only on one observation i, whereas the former problem yields the inequalities $\mathrm{E}[Y_i - Y_j|X_i, X_j, X_i > X_j] \geqslant 0$, each of which depends on the pair of observations (i, j). The basic ideas used in this article on constructing adaptive and rate optimal tests for the former problem, however, can be traced back to the results in Armstrong (2014), Armstrong and Chan (2016), and Chetverikov (2016) for the latter problem.

The rest of the article is organized as follows. Section 2 describes the general class of test statistics and gives several methods to simulate the critical value. Section 3 contains the main results under high-level conditions. Section 4 uses the results in Section 3 to construct an adaptive and rate optimal test under mild low-level conditions. Section 5 presents a small Monte Carlo simulation study. Section 6 concludes. All proofs are contained in the appendix.

2. TESTS

2.1. The Test Statistic

Let $Q(\cdot,\cdot)\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ be a non-negative and symmetric weighting function, so that $Q(x_1,x_2) = Q(x_2,x_1)$ and $Q(x_1,x_2) \geqslant 0$ for all $x_1, x_2 \in \mathbb{R}$, and let

$$b = \frac{1}{2}\sum_{1\leqslant i,j\leqslant n} (Y_i - Y_j)\mathrm{sign}(X_j - X_i)Q(X_i, X_j)$$

be a test function. Since $Q(X_i, X_j) \geqslant 0$ and $\mathrm{E}[Y_i|X_i] = f(X_i)$, it is easy to see that under $\mathcal{H}_0$, that is, when the function $f(\cdot)$ is nondecreasing, $\mathrm{E}[b] \leqslant 0$. On the other hand, if $\mathcal{H}_0$ is violated and there exist x_1 and x_2 on the support of X such that $x_1 < x_2$ but $f(x_1) > f(x_2)$, there exists a function $Q(\cdot,\cdot)$ such that $\mathrm{E}[b] > 0$ if $f(\cdot)$ is smooth. Therefore, b can be used to form a test statistic if there is an effective mechanism to find an appropriate weighting function $Q(\cdot,\cdot)$. For this purpose, I will use the adaptive testing approach developed in the statistics literature.

The idea behind the adaptive testing approach is to choose $Q(\cdot,\cdot)$ from a large set of potentially useful weighting functions that maximizes the studentized version of b. Formally, let $\mathcal{S}_n$ be some general set that depends on n and is (implicitly) allowed to depend on $\{X_i\}_{1\leqslant i\leqslant n}$, and for $s \in \mathcal{S}_n$, let $Q(\cdot,\cdot,s)\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ be some weighting function such that $Q(x_1,x_2,s) = Q(x_2,x_1,s)$ and $Q(x_1,x_2,s) \geqslant 0$ for all $x_1, x_2 \in \mathbb{R}$. The functions $Q(\cdot,\cdot,s)$ are also (implicitly) allowed to depend on $\{X_i\}_{1\leqslant i\leqslant n}$. In addition, let

$$b(s) = \frac{1}{2}\sum_{1\leqslant i,j\leqslant n} (Y_i - Y_j)\mathrm{sign}(X_j - X_i)Q(X_i, X_j, s) \tag{2}$$

be a test function. To derive the variance of $b(s)$, note that $b(s)$ can be equivalently rewritten as

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

$$b(s) = \sum_{i=1}^{n} Y_i \left(\sum_{1 \leqslant j \leqslant n} \operatorname{sign}(X_j - X_i) Q(X_i, X_j, s) \right). \tag{3}$$

Hence, conditional on $\{X_i\}_{1 \leqslant i \leqslant n}$, the variance of $b(s)$ is given by

$$V(s) = \sum_{1 \leqslant i \leqslant n} \sigma_i^2 \left(\sum_{1 \leqslant j \leqslant n} \operatorname{sign}(X_j - X_i) Q(X_i, X_j, s) \right)^2, \tag{4}$$

where $\sigma_i = (\mathrm{E}[\varepsilon_i^2 | X_i])^{1/2}$ and $\varepsilon_i = Y_i - f(X_i)$. In general, σ_i's are unknown, and have to be estimated from the data. For all $i = 1, \ldots, n$, let $\widehat{\sigma}_i$ denote some estimator of σ_i. Available estimators are discussed later in this section. Then the estimated conditional variance of $b(s)$ is

$$\widehat{V}(s) = \sum_{1 \leqslant i \leqslant n} \widehat{\sigma}_i^2 \left(\sum_{1 \leqslant j \leqslant n} \operatorname{sign}(X_j - X_i) Q(X_i, X_j, s) \right)^2. \tag{5}$$

The general form of the test statistic that I consider in this article is

$$T = \max_{s \in \mathcal{S}_n} \frac{b(s)}{(\widehat{V}(s))^{1/2}}. \tag{6}$$

Large values of T indicate that the null hypothesis $\mathcal{H}_0$ is violated. Later in this section, I will provide methods for estimating (or bounding) quantiles of T under $\mathcal{H}_0$ and for choosing a critical value for the test based on the statistic T.

The set $\mathcal{S}_n$ determines adaptivity properties of the test, that is the ability of the test to detect many different deviations from $\mathcal{H}_0$. Indeed, each weighting function $Q(\cdot, \cdot, s)$ is useful for detecting some deviation, and so the larger is the set of weighting functions $\mathcal{S}_n$, the larger is the number of different deviations that can be detected, and the higher is adaptivity of the test. In this article, I allow for *exponentially* large (in the sample size n) sets $\mathcal{S}_n$. This implies that the researcher can choose a huge set of weighting functions, which allows her to detect a large set of different deviations from $\mathcal{H}_0$. The downside of the adaptivity, however, is that expanding the set $\mathcal{S}_n$ increases the critical value, and thus decreases the power of the test against those alternatives that can be detected by weighting functions already included in $\mathcal{S}_n$. Fortunately, the loss of power is relatively small. In particular, it follows from Lemmas A.3 and A.6 in the appendix that the critical values for the tests developed below are bounded from above by a slowly growing $C(\log p)^{1/2}$ for some constant $C > 0$ where $p = |\mathcal{S}_n|$, the number of elements in the set $\mathcal{S}_n$.

2.2. Typical Weighting Functions

Let me now describe typical weighting functions. Consider some compactly supported kernel function $K : \mathbb{R} \to \mathbb{R}$ satisfying $K(x) \geqslant 0$ for all $x \in \mathbb{R}$. For convenience, I will assume that the support of $K(\cdot)$ is $[-1, 1]$. In addition, let $s = (x, h)$ where x is a location point and h is a bandwidth value (smoothing parameter). Finally, define

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

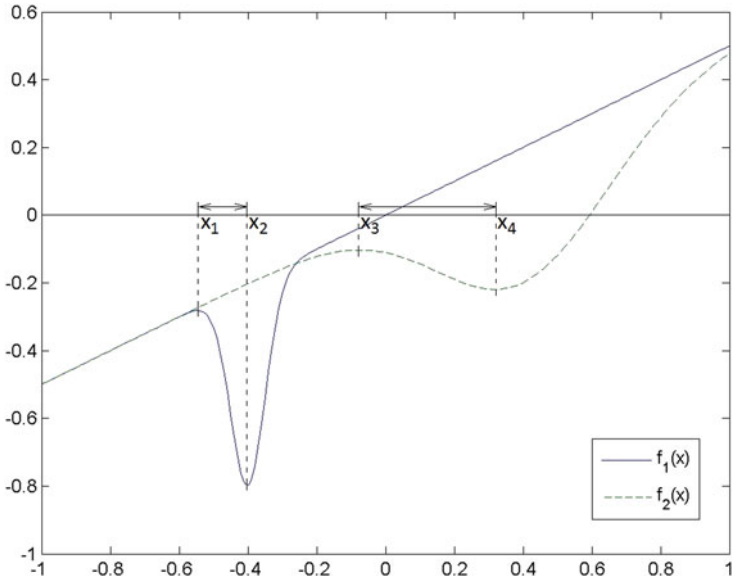

FIGURE 1. Regression functions illustrating different deviations from $\mathcal{H}_0$.

$$Q(x_1, x_2, s) = |x_1 - x_2|^k K\left(\frac{x_1 - x}{h}\right) K\left(\frac{x_2 - x}{h}\right) \tag{7}$$

for some $k \geqslant 0$. I refer to this Q as a *kernel weighting function*.[4]

Assume that a test is based on kernel weighting functions and $\mathcal{S}_n$ consists of pairs $s = (x, h)$ with many different values of x and h. Such a test would have good adaptivity properties. To see this, consider Figure 1 that plots two regression functions, $f_1(\cdot)$ and $f_2(\cdot)$. Both $f_1(\cdot)$ and $f_2(\cdot)$ violate $\mathcal{H}_0$ but locations where $\mathcal{H}_0$ is violated are different. In particular, $f_1(\cdot)$ violates $\mathcal{H}_0$ on the interval $[x_1, x_2]$ and $f_2(\cdot)$ violates $\mathcal{H}_0$ on the interval $[x_3, x_4]$. In addition, $f_1(\cdot)$ is relatively less smooth than $f_2(\cdot)$, and $[x_1, x_2]$ is shorter than $[x_3, x_4]$. To have good power against $f_1(\cdot)$, $\mathcal{S}_n$ should contain a pair (x, h) such that $[x - h, x + h] \subset [x_1, x_2]$. Indeed, if $[x - h, x + h]$ is not contained in $[x_1, x_2]$, then positive and negative values of the summand in b will cancel out yielding a low value of b. In particular, it should be the case that $x \in [x_1, x_2]$. Similarly, to have good power against $f_2(\cdot)$, $\mathcal{S}_n$ should contain a pair (x, h) such that $x \in [x_3, x_4]$. Therefore, using many different values of x yields a test that adapts to the location of the deviation from $\mathcal{H}_0$. This is spatial adaptivity. Furthermore, note that larger values of h yield higher signal-to-noise ratio. So, given that $[x_3, x_4]$ is longer than $[x_1, x_2]$, the optimal pair (x, h) to test against $f_2(\cdot)$ has a larger value of h than that used to test against $f_1(\cdot)$.

[4] It is possible to extend the definition of kernel weighting functions given in (7). Specifically, the term $|x_1 - x_2|^k$ in the definition can be replaced by general function $\bar{K}(x_1, x_2)$ satisfying $\bar{K}(x_1, x_2) \geqslant 0$ for all x_1 and x_2. I thank Joris Pinkse for this observation.

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

Therefore, using many different values of h results in adaptivity with respect to smoothness of the function, which, in turn, determines how fast its first derivative is varying and how long the interval of nonmonotonicity is.

If no ex ante information is available, I recommend using kernel weighting functions (7) with

$$\mathcal{S}_n = \Big\{(x,h)\colon x \in \{X_1,\dots,X_n\}, h \in H_n\Big\},$$

where

$$H_n = \Big\{h = h_{\max}u^l\colon h \geqslant h_{\min}, l = 0,1,2,\dots\Big\},$$

$$h_{\max} = \max_{1\leqslant i,j\leqslant n} |X_i - X_j|/2, \text{ and } h_{\min} = C_h h_{\max}(\log n/n)^{1/3}.$$

I also recommend setting $u = 0.5$, $C_h = 0.4$, and $k = 0$ or 1. I refer to this $\mathcal{S}_n$ as a *basic set* of weighting functions. This choice of parameters is consistent with the theory presented in this article and has worked well in simulations. Moreover, the basic set of weighting functions yields an adaptive and rate optimal test; see Section 4. The constant C_h is selected so that each test function $b(s)$ uses approximately at least 15 observations when $n = 100$ and X is distributed uniformly on some interval.

If some ex ante information is available, the general framework considered here gives the researcher a lot of flexibility to incorporate this information. In particular, if the researcher expects that the function $f(\cdot)$ is rather smooth, she can restrict the set $\mathcal{S}_n$ by considering only pairs (x,h) with large values of h since in this case deviations from $\mathcal{H}_0$, if present, are more likely to happen on long intervals. Moreover, if the smoothness of the function $f(\cdot)$ is known, one can find an optimal value of the smoothing parameter $\tilde{h} = \tilde{h}_n$ corresponding to this level of smoothness, and then consider kernel weighting functions with this particular choice of the bandwidth value, that is $\mathcal{S}_n = \{(x,h)\colon x \in \{X_1,\dots,X_n\}, h = \tilde{h}\}$. Furthermore, if nonmonotonicity is expected at one particular point $\tilde{x}$, one can consider kernel weighting functions with $\mathcal{S}_n = \{(x,h)\colon x = \tilde{x}, h = \tilde{h}\}$ or $\mathcal{S}_n = \{(x,h)\colon x = \tilde{x}, h \in H_n\}$ depending on whether the smoothness of $f(\cdot)$ is known or not. More broadly, if nonmonotonicity is expected on some interval $\bar{\mathcal{X}}$, one can use kernel weighting functions with $\mathcal{S}_n = \{(x,h)\colon x \in \{X_1,\dots,X_n\} \cap \bar{\mathcal{X}}, h \in \tilde{h}\}$ or $\mathcal{S}_n = \{(x,h)\colon x \in \{X_1,\dots,X_n\} \cap \bar{\mathcal{X}}, h \in H_n\}$ again depending on whether the smoothness of $f(\cdot)$ is known or not. Note that all these modifications will increase the power of the test because smaller sets $\mathcal{S}_n$ yield lower critical values.

Another interesting choice of weighting functions is

$$Q(x_1,x_2,s) = \sum_{1\leqslant r\leqslant m} |x_1 - x_2|^k K\left(\frac{x_1 - x^r}{h}\right) K\left(\frac{x_2 - x^r}{h}\right),$$

where $s = (x^1,\dots,x^m,h)$. These weighting functions are useful if the researcher expects multiple deviations from $\mathcal{H}_0$.

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

2.3. Comparison with Other Known Tests

I will now show that the general framework described above includes the Hall and Heckman's (HH) test statistic and a slightly modified version of the Ghosal, Sen, and van der Vaart's (GSV) test statistic as special cases that correspond to different values of k in the definition of the kernel weighting functions (7).

First, consider the GSV test. This test is based on the test functions

$$b(s) = \frac{1}{2} \sum_{1 \leqslant i,j \leqslant n} \operatorname{sign}(Y_i - Y_j)\operatorname{sign}(X_j - X_i) K\left(\frac{X_i - x}{h}\right) K\left(\frac{X_j - x}{h}\right),$$

whereas setting $k = 0$ in (7) yields

$$b(s) = \frac{1}{2} \sum_{1 \leqslant i,j \leqslant n} (Y_i - Y_j)\operatorname{sign}(X_j - X_i) \left(\frac{X_i - x}{h}\right) K\left(\frac{X_j - x}{h}\right). \tag{8}$$

Hence, the only difference between the GSV test and my test is that I use the term $(Y_i - Y_j)$ whereas they use $\operatorname{sign}(Y_i - Y_j)$. It will be shown in the next section that my tests are consistent. On the other hand, one can see that the GSV test is not consistent under the presence of conditional heteroscedasticity. Indeed, assume that X_i is supported on $[0, 1]$, that $f(X_i) = -X_i$, and that ε_i is $-2X_i$ or $2X_i$ with equal probabilities. Then $(Y_i - Y_j)(X_j - X_i) > 0$ if and only if $(\varepsilon_i - \varepsilon_j)(X_j - X_i) > 0$, and so the probability of rejecting $\mathcal{H}_0$ for the GSV test is numerically equal to that in the model with $f(\cdot) \equiv 0$. But the latter probability does not exceed the size of the test. This implies that the GSV test is not consistent since it maintains the required size asymptotically.[5] Moreover, the GSV test is nonadaptive with respect to the smoothness of the function $f(\cdot)$.

Next, consider the HH test. The idea of this test is to make use of the local linear estimates of the slope of the function $f(\cdot)$. Using well-known formulas for the OLS regression, one can show that the slope estimate of the function $f(\cdot)$ given the data $\{X_i, Y_i\}_{s_1 < i \leqslant s_2}$ with $s_1 < s_2$ where $\{X_i\}_{1 \leqslant i \leqslant n}$ is an increasing sequence is given by

$$b(s) = \frac{\sum_{s_1 < i \leqslant s_2} Y_i \sum_{s_1 < j \leqslant s_2} (X_i - X_j)}{(s_2 - s_1)\sum_{s_1 < i \leqslant s_2} X_i^2 - (\sum_{s_1 < i \leqslant s_2} X_i)^2}, \tag{9}$$

where $s = (s_1, s_2)$. Note that the denominator of (9) depends only on X_i's, and so it disappears after studentization. In addition, simple rearrangements show that the numerator in (9) is up to the sign equal to

$$\frac{1}{2} \sum_{1 \leqslant i,j \leqslant n} (Y_i - Y_j)(X_j - X_i) 1\{x - h \leqslant X_i \leqslant x + h\} 1\{x - h \leqslant X_j \leqslant x + h\} \tag{10}$$

[5] The same conclusion on inconsistency also applies to the test of Gijbels et al. (2000).

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

for some x and h. On the other hand, setting $k = 1$ in (7) yields

$$b(s) = \frac{1}{2} \sum_{1 \leqslant i,j \leqslant n} (Y_i - Y_j)(X_j - X_i) K\left(\frac{X_i - x}{h}\right) K\left(\frac{X_j - x}{h}\right). \tag{11}$$

Hence, the expression in (10) is proportional to that on the right-hand side in (11) with $K(\cdot) = 1\{-1 \leqslant \cdot \leqslant +1\}$, and so the HH test statistic is a special case of those studied in this article. Note also that the HH test statistic maximizes the studentized version of $b(s)$ over $s_1 < s_2$, and so it corresponds to the basic set of weighting functions with the given kernel function $K(\cdot)$ and leads to an adaptive and rate optimal test.

2.4. Estimating σ_i's

In practice, the σ_i's are usually unknown, and have to be estimated from the data. In this subsection, I explain how this estimation can be carried out.

First, if the regression model (1) is homoscedastic, so that $\sigma_i = \sigma$ for all $i = 1, \ldots, n$ and some σ, one can use the estimator of Rice (1984):

$$\widehat{\sigma} = \left(\frac{1}{2n} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2\right)^{1/2}, \tag{12}$$

where it is assumed that the observations $\{X_i, Y_i\}_{1 \leqslant i \leqslant n}$ are arranged so that $X_i \leqslant X_j$ whenever $i \leqslant j$. This estimator is $\sqrt{n}$-consistent for σ under mild regularity conditions as long as $f(\cdot)$ is piecewise Lipschitz-continuous.

Second, if the regression model (1) is heteroscedastic, so that σ_i varies over $i = 1, \ldots, n$, one can use a local version of the Rice estimator. To define the estimator, let $b_n > 0$ be some bandwidth value. Also, for $i = 1, \ldots, n$, let

$$J(i) = \left\{j = 1, \ldots, n \colon |X_j - X_i| \leqslant b_n\right\},$$

and let $|J(i)|$ denote the number of elements in $J(i)$. Then the local Rice estimator is

$$\widehat{\sigma}_i = \left(\frac{1}{2|J(i)|} \sum_{j \in J(i) \colon j+1 \in J(i)} (Y_{j+1} - Y_j)^2\right)^{1/2}, \quad i = 1, \ldots, n. \tag{13}$$

The intuition behind this estimator is as follows. Note that if b_n is small enough, X_{j+1} is close to X_j for all $j \in J(i)$ such that $j + 1 \in J(i)$. So, if the function $f(\cdot)$ is continuous,

$$Y_{j+1} - Y_j = f(X_{j+1}) - f(X_j) + \varepsilon_{j+1} - \varepsilon_j \approx \varepsilon_{j+1} - \varepsilon_j,$$

so that

$$\mathrm{E}\left[(Y_{j+1} - Y_j)^2 | \{X_i\}_{1 \leqslant i \leqslant n}\right] \approx \sigma_{j+1}^2 + \sigma_j^2$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

since ε_{j+1} is independent of ε_j and $\mathrm{E}[\varepsilon_j|\{X_i\}_{1\leqslant i\leqslant n}]=0$. In addition, if b_n is small enough and the function $\mathrm{E}[\varepsilon^2|X=\cdot]$ is continuous, $\sigma^2_{j+1}+\sigma^2_j\approx 2\sigma^2_i$ since $|X_{j+1}-X_i|\leqslant b_n$ and $|X_j-X_i|\leqslant b_n$. Hence, if $|J(i)|$ is large enough, which happens with high probability if b_n is not too small, $\widehat{\sigma}^2_i$ is close to σ^2_i by the law of large numbers. The formal properties of this estimator will be given in Section 4. Other available estimators are presented, for example, in Muller and Stadtmuller (1987), Fan and Yao (1998), Horowitz and Spokoiny (2001), Hardle and Tsybakov (2007), and Cai and Wang (2008).

2.5. Simulating the Critical Value

In this subsection, I provide three methods for estimating (or bounding) quantiles of the null distribution of the test statistic T. These are plug-in, one-step, and step-down methods. All of these methods are based on the procedure known as the Wild bootstrap. The Wild bootstrap was introduced in Wu (1986) and used, among many others, by Liu (1988), Mammen (1993), Hardle and Mammen (1993), Horowitz and Spokoiny (2001), Chetverikov (2016), and Chernozhukov, Chetverikov, and Kato (2013, 2016a, 2017). The three methods are arranged in terms of increasing power and computational complexity. The validity of all three methods is established in Theorem 3.1 below. To describe the methods, let $\{\epsilon_i\}_{1\leqslant i\leqslant n}$ be i.i.d. $N(0,1)$ random variables that are independent of the data.

2.6. Plug-in Approach

Suppose that we are interested in obtaining a test of level α. Throughout the rest of the article, and without further notice, I assume that $\alpha\in(0,1/2)$. The plug-in approach is based on two observations. First, under $\mathcal{H}_0$, for all $s\in\mathcal{S}_n$,

$$b(s)=\frac{1}{2}\sum_{1\leqslant i,j\leqslant n}(f(X_i)-f(X_j)+\varepsilon_i-\varepsilon_j)\mathrm{sign}(X_j-X_i)Q(X_i,X_j,s) \tag{14}$$

$$\leqslant\frac{1}{2}\sum_{1\leqslant i,j\leqslant n}(\varepsilon_i-\varepsilon_j)\mathrm{sign}(X_j-X_i)Q(X_i,X_j,s) \tag{15}$$

since $Q(X_i,X_j,s)\geqslant 0$ and $f(X_i)\geqslant f(X_j)$ whenever $X_i\geqslant X_j$ under $\mathcal{H}_0$. Hence, the $(1-\alpha)$ quantile of T is bounded from above by the $(1-\alpha)$ quantile of T in the model with $f(\cdot)\equiv 0$, which is the least favorable model under $\mathcal{H}_0$. Second, I will show that conditional on $\{X_i\}_{1\leqslant i\leqslant n}$, the distribution of T asymptotically depends on the distribution of noise $\{\varepsilon_i\}_{1\leqslant i\leqslant n}$ only through second moments $\{\sigma^2_i\}_{1\leqslant i\leqslant n}$; see Lemma A.5 in the appendix. These two observations suggest that the critical value for the test can be obtained by simulating the $(1-\alpha)$ quantile of the conditional distribution of T given $\{X_i\}_{1\leqslant i\leqslant n}$ in the model with $f(\cdot)\equiv 0$, fixed $\{X_i\}_{1\leqslant i\leqslant n}$, and Gaussian noise, so that $\varepsilon_i\sim N(0,\widehat{\sigma}^2_i)$ for all $i=1,\ldots,n$. More precisely, I define the plug-in critical value $c^{PI}_{1-\alpha}$ as the $(1-\alpha)$ quantile of the

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

conditional distribution of the bootstrap test statistic

$$T^{\mathcal{S}_n} = \max_{s \in \mathcal{S}_n} \frac{\frac{1}{2}\sum_{1 \leqslant i,j \leqslant n} (\widehat{\sigma}_i \epsilon_i - \widehat{\sigma}_j \epsilon_j) \text{sign}(X_j - X_i) Q(X_i, X_j, s)}{(\widehat{V}(s))^{1/2}}$$

given the data $\{X_i, Y_i\}_{1 \leqslant i \leqslant n}$.

2.7. One-Step Approach

The test with the plug-in critical value is computationally simple. It has, however, poor power properties. Indeed, the distribution of T in general depends on $f(\cdot)$ but the plug-in approach is based on the least favorable regression function $f(\cdot) \equiv 0$, and so it is too conservative when $f(\cdot)$ has strictly increasing parts. To (partially) fix this problem, let $\{\gamma_n\}_{n \geqslant 1}$ be a sequence of positive numbers converging to zero, and let $c^{PI}_{1-\gamma_n}$ be the $(1-\gamma_n)$ plug-in critical value. In what follows, I refer to the number γ_n as a threshold probability. In addition, denote

$$\mathcal{S}_n^{OS} = \left\{ s \in \mathcal{S}_n : b(s)/(\widehat{V}(s))^{1/2} > -2c^{PI}_{1-\gamma_n} \right\}.$$

I define the one-step critical value $c^{OS}_{1-\alpha}$ as the $(1-\alpha)$ quantile of the conditional distribution of the bootstrap test statistic

$$T^{\mathcal{S}_n^{OS}} = \max_{s \in \mathcal{S}_n^{OS}} \frac{\frac{1}{2}\sum_{1 \leqslant i,j \leqslant n} (\widehat{\sigma}_i \epsilon_i - \widehat{\sigma}_j \epsilon_j) \text{sign}(X_j - X_i) Q(X_i, X_j, s)}{(\widehat{V}(s))^{1/2}}$$

given the data $\{X_i, Y_i\}_{1 \leqslant i \leqslant n}$.[6] Intuitively, the one-step critical value $c^{OS}_{1-\alpha}$ works because the weighting functions corresponding to elements of the set $\mathcal{S}_n \backslash \mathcal{S}_n^{OS}$ have an asymptotically negligible influence on the distribution of T under $\mathcal{H}_0$. Indeed, I will show that the probability that at least one element s of $\mathcal{S}_n$ such that

$$\frac{\frac{1}{2}\sum_{1 \leqslant i,j \leqslant n} (f(X_i) - f(X_j)) \text{sign}(X_j - X_i) Q(X_i, X_j, s)}{(\widehat{V}(s))^{1/2}} > -c^{PI}_{1-\gamma_n} \tag{16}$$

belongs to the set $\mathcal{S}_n \backslash \mathcal{S}_n^{OS}$ is at most $\gamma_n + o(1)$; see Lemma A.7 in the appendix. On the other hand, the probability that at least one element s of $\mathcal{S}_n$ such that (16) does not hold for this element gives $b(s)/(\widehat{V}(s))^{1/2} > 0$ is again at most $\gamma_n + o(1)$; see Lemma A.8 in the appendix. Since γ_n converges to zero, this suggests that the critical value can be simulated using only elements of $\mathcal{S}_n^{OS}$. In practice, one can set γ_n as a small fraction of α. For example, the Monte Carlo simulations presented in this article use $\gamma_n = 0.01$ with $\alpha = 0.1$.[7]

[6] If $\mathcal{S}_n^{OS}$ turns out to be empty, set $c^{OS}_{1-\alpha} = +\infty$.

[7] More formally, it is shown in the proof of Theorem 3.1 that the probability of rejecting $\mathcal{H}_0$ under $\mathcal{H}_0$ in large samples is bounded from above by $\alpha + 2\gamma_n$. This suggests that if the researcher does not agree to tolerate small size distortions, she can use the test with level $\tilde{\alpha} = \alpha - 2\gamma_n$ instead. On the other hand, I note that $\alpha + 2\gamma_n$ is only an *upper* bound on the probability of rejecting $\mathcal{H}_0$, and in many cases the true probability of rejecting $\mathcal{H}_0$ is smaller than $\alpha + 2\gamma_n$.

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

2.8. Step-Down Approach

The one-step approach, as the name suggests, uses only one step to drop those elements of $\mathcal{S}_n$ that have negligible influence on the distribution of T. It turns out that this step can be iterated using the step-down procedure and yielding second-order improvements in the power. The step-down procedures were developed in the literature on multiple hypothesis testing; see, in particular, Holm (1979), Romano and Wolf (2005a,b), and Romano and Shaikh (2010). See also Lehmann and Romano (2005) for a textbook introduction. The use of the step-down method in this article, however, is rather different.

To explain the step-down approach, let me define the sequences $\{c^l_{1-\gamma_n}\}_{l\geqslant 1}$ and $\{\mathcal{S}^l_n\}_{l\geqslant 1}$. Set $c^1_{1-\gamma_n} = c^{OS}_{1-\gamma_n}$ and $\mathcal{S}^1_n = \mathcal{S}^{OS}_n$. Then for $l > 1$, let $c^l_{1-\gamma_n}$ be the $(1-\gamma_n)$ quantile of the conditional distribution of

$$T^{\mathcal{S}^l_n} = \max_{s\in\mathcal{S}^l_n} \frac{\frac{1}{2}\sum_{1\leqslant i,j\leqslant n}(\widehat{\sigma}_i\epsilon_i - \widehat{\sigma}_j\epsilon_j)\text{sign}(X_j - X_i)Q(X_i, X_j, s)}{(\widehat{V}(s))^{1/2}}$$

given the data $\{X_i, Y_i\}_{1\leqslant i\leqslant n}$ where

$$\mathcal{S}^l_n = \left\{s \in \mathcal{S}_n : b(s)/(\widehat{V}(s))^{1/2} > -c^{PI}_{1-\gamma_n} - c^{l-1}_{1-\gamma_n}\right\}.$$

It is easy to see that $(c^l_{1-\gamma_n})_{l=1}^{\infty}$ is a decreasing sequence, and so $\mathcal{S}^l_n \supseteq \mathcal{S}^{l+1}_n$ for all $l \geqslant 1$. Since $\mathcal{S}^1_n$ is a finite set, there exists $l(0) \geqslant 1$ such that $\mathcal{S}^l_n = \mathcal{S}^{l+1}_n$ for all $l \geqslant l(0)$. Let $\mathcal{S}^{SD}_n = \mathcal{S}^{l(0)}_n$. I define the step-down critical value $c^{SD}_{1-\alpha}$ as the $(1-\alpha)$ quantile of the conditional distribution of the bootstrap test statistic $T^{\mathcal{S}^{l(0)}_n}$ given the data $\{X_i, Y_i\}_{1\leqslant i\leqslant n}$.

Note that $\mathcal{S}^{SD}_n \subset \mathcal{S}^{OS}_n \subset \mathcal{S}_n$, and so $c^{SD}_{1-\alpha} \leqslant c^{OS}_{1-\alpha} \leqslant c^{PI}_{1-\alpha}$. This explains that the three methods for simulating the critical values are arranged in terms of increasing power.

3. THEORY UNDER HIGH-LEVEL CONDITIONS

In this section, I present the main results of the article on the size and the power properties of the tests. Since some of the conditions used to derive the results are high-level, I will verify those conditions for the basic set of weighting functions and the local Rice estimator in the next section. I will also use the results from this section to obtain an adaptive and rate optimal test of monotonicity under low-level conditions in the next section.

Let c_1, C_1, c_2, C_2, c_3, C_3, β, and L be some constants such that $0 < c_1 \leqslant C_1$, $c_2 < C_2$, $0 < c_3 \leqslant C_3$, $0 < \beta \leqslant 1$, and $L > 0$. Also, let $\sigma(\cdot) = (\text{E}[\varepsilon^2|X = \cdot])^{1/2}$ be the heteroscedasticity function. Moreover, for any differentiable function $g : \mathbb{R} \to \mathbb{R}$, let $g'(\cdot)$ denote its derivative. Throughout the article, I will use the following low-level assumptions.

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

Assumption L 1 (Noise moments)**.** The random variable ε is such that $(\mathrm{E}[|\varepsilon|^4|X])^{1/4} \leqslant C_1$ and $\sigma(X) \geqslant c_1$ almost surely.

Assumption L2 (Distribution of X)**.** The support of X is $\mathcal{X} = [c_2, C_2]$, and the distribution of X is absolutely continuous with respect to the Lebesgue measure on $\mathcal{X}$ with the pdf bounded from below by c_3 and from above by C_3 on $\mathcal{X}$.

Assumption L3 (Smoothness)**.** The regression function $f(\cdot)$ is continuously differentiable and is such that $|f'(x)| \leqslant L$ for all $x \in \mathcal{X}$ and $|f'(x_2) - f'(x_1)| \leqslant L|x_2 - x_1|^\beta$ for all $x_1, x_2 \in \mathcal{X}$. In addition, the heteroscedasticity function $\sigma(\cdot)$ is such that $|\sigma(x_2) - \sigma(x_1)| \leqslant L|x_2 - x_1|$ for all $x_1, x_2 \in \mathcal{X}$.

These are mild assumptions on the moments of the noise variable ε, on the distribution of X, and on the smoothness of the regression and heteroscedasticity functions $f(\cdot)$ and $\sigma(\cdot)$. The condition that $\sigma(X) \geqslant c_1$ almost surely imposed in Assumption L1 precludes the existence of super-efficient estimators. Assumption L2 means that for any $x_1, x_2 \in \mathcal{X}$ with $x_1 < x_2$,

$$c_3(x_2 - x_1) \leqslant \mathrm{P}(x_1 \leqslant X \leqslant x_2) \leqslant C_3(x_2 - x_1).$$

This assumption precludes discrete X's, but I note that when X is discrete, the function $f(\cdot)$ can be estimated with the $\sqrt{n}$-rate of consistency for all points on the support of X and the problem of testing monotonicity of $f(\cdot)$ reduces to a standard parametric testing problem. Assumption L3 requires, among other things, that $f(\cdot)$ is continuously differentiable with the derivative being Lipschitz-continuous with Lipschitz constant L and Lipschitz order β.

Furthermore, let c_4, C_4, c_5, and κ be some constants such that $0 < c_4 \leqslant C_4$, $c_5 > 0$, $0 < \kappa \leqslant 1/2$, and define the following sensitivity parameter:

$$A_n = \max_{s \in \mathcal{S}_n} \max_{1 \leqslant i \leqslant n} \left| \sum_{1 \leqslant j \leqslant n} \frac{\mathrm{sign}(X_j - X_i) Q(X_i, X_j, s)}{(V(s))^{1/2}} \right|. \tag{17}$$

This (random-valued) parameter provides an upper bound on the weights each observation i may have in the studentized test functions $b(s)/(V(s))^{1/2}$; see (3). Moreover, let $p = |\mathcal{S}_n|$ denote the number of elements in the set $\mathcal{S}_n$, and let $h_n = (\log p/n)^{1/(2\beta+3)}$. To derive the results, I will also use the following high-level assumptions in addition to low-level Assumptions L1–L3.

Assumption H1 (Noise variance estimators)**.** The estimators $\widehat{\sigma}_i$, $i = 1, \ldots, n$, are such that $\max_{1 \leqslant i \leqslant n} |\widehat{\sigma}_i - \sigma_i| = o_p(n^{-\kappa})$.

Assumption H2 (Test function variance estimators)**.** The estimators $\widehat{V}(s)$, $s \in \mathcal{S}_n$, are such that $\max_{s \in \mathcal{S}_n} |(\widehat{V}(s)/V(s))^{1/2} - 1| = o_p(n^{-\kappa})$.

Assumption H3 (Growth conditions)**.** The sensitivity parameter A_n is such that $nA_n^4 \log^7(pn) = o_p(1)$ and the number of weighting functions p is such that $\log p/n^\kappa = o(1)$.

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

Assumption H4 (Weighting functions)**.** With probability $1 - o(1)$, for any interval $[x_1, x_2] \subset \mathcal{X}$ with $x_2 - x_1 \geqslant h_n$, there exists $s \in \mathcal{S}_n$ such that (i) $Q(y_1, y_2, s) = 0$ for all $(y_1, y_2) \notin [x_1, x_2]^2$, (ii) $Q(y_1, y_2, s) \leqslant C_4(x_2 - x_1)^k$ for all $(y_1, y_2) \in [x_1, x_2]^2$, and (iii) there exist nonintersecting subintervals $[x_{l1}, x_{r1}]$ and $[x_{l2}, x_{r2}]$ of $[x_1, x_2]$ such that $x_{r2} \geqslant x_{l2} + c_5(x_2 - x_1) \geqslant x_{r1} + 2c_5(x_2 - x_1) \geqslant x_{l1} + 3c_5(x_2 - x_1)$ and $Q(y_1, y_2, s) \geqslant c_4(x_2 - x_1)^k$ for all $(y_1, y_2) \in [x_{l1}, x_{r1}] \times [x_{l2}, x_{r2}]$.

These are high-level assumptions on the rate of consistency of the estimators $\widehat{\sigma}_i$, on the rate of consistency of the estimators $\widehat{V}(s)$, on the sensitivity parameter A_n, and on the weighting functions. In the next section, I will verify these assumptions for the basic set of weighting functions and the local Rice estimator. Note that Assumption H3 includes p only through $\log p$, and so it allows an exponentially large (in the sample size n) number of weighting functions. Also, note that it is shown in the proof of Theorem 4.2 in Appendix A.2 that when the kernel weighting functions with bandwidth values $h \geqslant h_{\min}$ are used, there exists a constant $C > 0$ such that the sensitivity parameter A_n with probability $1 - o(1)$ satisfies the bound $A_n \leqslant C/(nh_{\min})^{1/2}$. Hence, in this case, the condition that $nA_n^4 \log^7(pn) = o_p(1)$ holds as long as $n^{1-\epsilon} h_{\min}^2 \geqslant c$ and $\log p \leqslant C \log n$ for all $n \geqslant 1$ with arbitrarily small constants $c, \epsilon > 0$ and arbitrarily large constant $C > 0$.

Next, to state the main results on the size and the power properties of the tests, let M be a model given by the regression function $f(\cdot)$ and the joint distribution of the pair (X, ε) such that $\mathrm{E}[\varepsilon | X] = 0$ almost surely. For the model M, I assume that the dependent variable Y is generated from the pair (X, ε) and the regression function $f(\cdot)$ according to (1), so that $Y = f(X) + \varepsilon$. Also, let $\mathcal{M}_L$ denote the set of all models M that satisfy Assumptions L1, L2, and L3 (with the same constants c_1, C_1, c_2, C_2, c_3, C_3, β, and L). Since different models $M \in \mathcal{M}_{\mathcal{L}}$ may have different regression functions $f(\cdot)$, I will sometimes index the regression function $f(\cdot)$ by the model M: $f(\cdot) = f_M(\cdot)$. Furthermore, let $\mathcal{M}_{\mathcal{H}}$ denote a set of models $M \in \mathcal{M}_{\mathcal{L}}$ such that Assumptions H1–H4 hold uniformly over this set.[8] In the next section, I will show that if the basic set of weighting functions and the local Rice estimator are used, one can take $\mathcal{M}_{\mathcal{H}} = \mathcal{M}_{\mathcal{L}}$. For $M \in \mathcal{M}_{\mathcal{L}}$, let $\mathrm{P}_M(\cdot)$ denote the probability measure generated by the model M. The following theorem shows that the tests developed in this article control asymptotic size uniformly over the class $\mathcal{M}_{\mathcal{H}}$.

THEOREM 3.1 (Size properties of the tests). *Let $P = PI$, OS, or SD and let $\mathcal{M}_{0,\mathcal{H}}$ denote the set of all models $M \in \mathcal{M}_{\mathcal{H}}$ satisfying $\mathcal{H}_0$. Then*

$$\inf_{M \in \mathcal{M}_{0,\mathcal{H}}} \mathrm{P}_M(T \leqslant c_{1-\alpha}^P) \geqslant 1 - \alpha + o(1)$$

as $n \to \infty$. In addition, let $\mathcal{M}_{00,\mathcal{H}}$ denote the set of all models $M \in \mathcal{M}_{0,\mathcal{H}}$ such

[8] Assumptions H1, H2, and H3 contain statements of the form $Z = o_p(n^{-\kappa})$ for some random variable Z and some constant $\kappa > 0$. I say that these assumptions hold uniformly over a set of models if for any $C > 0$, $\mathrm{P}(|Z| > Cn^{-\kappa}) = o(1)$ uniformly over this set.

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

that $f_M(\cdot) \equiv C$ *for some constant* C. *Then*

$$\sup_{M \in \mathcal{M}_{00,\mathcal{H}}} \left| \mathrm{P}_M(T \leqslant c^P_{1-\alpha}) - (1-\alpha) \right| \to 0$$

as $n \to \infty$.

Comment 3.1. (*i*) This theorem states that the wild bootstrap combined with the selection procedures developed in this article yields valid critical values in the sense that the resulting tests control asymptotic size. Moreover, critical values are valid uniformly over the class of models $\mathcal{M}_{0,\mathcal{H}}$, and, in addition, the tests are nonconservative in the sense that their level converges to the nominal level α.
(*ii*) The proof technique used in this theorem is based on *finite sample* approximations that are built on the results of Chernozhukov, Chetverikov, and Kato (2013, 2015). In particular, the validity of the bootstrap is established *without* referring to the asymptotic distribution of the test statistic. This is important because, for example, when the test is based on the basic set of weighting functions, the asymptotic distribution of the test statistic is unknown. Moreover, it is not even clear whether this distribution exists.
(*iii*) The standard techniques from the empirical process theory as presented, for example, in van der Vaart and Wellner (1996) are not sufficient to prove Theorem 3.1. The problem is that it is typically impossible to embed the process $\{b(s)/(V(s))^{1/2}\}_{s \in \mathcal{S}_n}$ into an asymptotically equicontinuous process since, for example, when the basic set of weighting functions is used, so that $s = (x,h)$, the random variables $b(x_1,h)/(V(x_1,h))^{1/2}$ and $b(x_2,h)/(V(x_2,h))^{1/2}$ for fixed $x_1 < x_2$ become asymptotically independent as $h \to 0$.
(*iv*) Note that T asymptotically has a form of U-statistic. The analysis of such statistics typically requires a preliminary Hoeffding projection. An advantage of the approximation method used in this article is that it applies directly to the test statistic with no need for the Hoeffding projection, which greatly simplifies the derivations. □

Next, I present power properties of the tests. The following theorem establishes consistency of the tests against fixed alternatives in the class $\mathcal{M}_{\mathcal{H}}$.

THEOREM 3.2 (Consistency against fixed alternatives). *Let* $P = PI$, OS, *or* SD. *Then for any model* $M \in \mathcal{M}_{\mathcal{H}}$ *such that there exist* $x_1, x_2 \in \mathcal{X}$ *with* $x_1 < x_2$ *but* $f_M(x_1) > f_M(x_2)$ ($\mathcal{H}_0$ *is false*),

$$\mathrm{P}_M(T \leqslant c^P_{1-\alpha}) \to 0 \text{ as } n \to \infty$$

as $n \to \infty$.

Furthermore, to derive the rate of consistency of the tests against local one-dimensional alternatives, consider any $M \in \mathcal{M}_{\mathcal{H}}$ such that there exist $x_1, x_2 \in \mathcal{X}$ with $x_1 < x_2$ but $f_M(x_1) > f_M(x_2)$ ($\mathcal{H}_0$ is false), and let $\{M_n\}_{n \geqslant 1}$ be a sequence of models such that for all $n \geqslant 1$, the joint distribution of the pair (X, ε) in M_n

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

coincide with that in M but the regression function $f_{M_n}(\cdot)$ in M_n has the following form: $f_{M_n}(\cdot) = l_n f_M(\cdot)$, where $\{l_n\}_{n\geqslant 1}$ is some sequence of positive numbers converging to zero. Note that all the models M_n are under $\mathcal{H}_a$ but they approach $\mathcal{H}_0$ as n increases.

THEOREM 3.3 (Rate of consistency against local one-dimensional alternatives). *Let $P = PI$, OS, or SD. Suppose that the sequence $\{M_n\}_{n\geqslant 1}$ defined above is such that $M_n \in \mathcal{M}_{\mathcal{H}}$ for all $n \geqslant 1$. Then*

$$\mathrm{P}_{M_n}(T \leqslant c^P_{1-\alpha}) \to 0$$

as $n \to \infty$ as long as $(\log p/n)^{1/2} = o(l_n)$.

Finally, to derive the rate of uniform consistency of the tests against alternatives in the class $\mathcal{M}_{\mathcal{H}}$, let $\{l_n\}_{n\geqslant 1}$ be a sequence of positive numbers converging to zero, and let $\mathcal{M}_{\mathcal{H},l_n}$ be the set of all models $M \in \mathcal{M}_{\mathcal{H}}$ such that $\inf_{x\in\mathcal{X}} f'_M(x) < -l_n$.

THEOREM 3.4 (Rate of uniform consistency against smooth alternatives). *Let $P = PI$, OS, or SD. Then*

$$\sup_{M\in\mathcal{M}_{\mathcal{H},l_n}} \mathrm{P}_M(T \leqslant c^P_{1-\alpha}) \to 0$$

as $n \to \infty$ as long as $(\log p/n)^{\beta/(2\beta+3)} = o(l_n)$.

To conclude this section, I present a theorem that gives a lower bound on the possible rates of uniform consistency against alternatives in the class $\mathcal{M}_{\mathcal{L}}$ so that no test that maintains asymptotic size can have a faster rate of uniform consistency. Let $\psi = \psi(\{X_i, Y_i\}_{1\leqslant i\leqslant n})$ be a generic test where the function $\psi(\{X_i, Y_i\}_{1\leqslant i\leqslant n})$ gives the probability of rejecting $\mathcal{H}_0$ upon observing the data $\{X_i, Y_i\}_{1\leqslant i\leqslant n}$. Also, let $\mathcal{M}_{0,\mathcal{L}}$ denote the set of all models $M \in \mathcal{M}_{\mathcal{L}}$ satisfying $\mathcal{H}_0$. Moreover, for $M \in \mathcal{M}_{\mathcal{L}}$, let $\mathrm{E}_M[\cdot]$ denote the expectation under the probability measure generated by the model M.

THEOREM 3.5 (Lower bound on possible rates of uniform consistency). *For any test ψ such that*

$$\sup_{M\in\mathcal{M}_{0,\mathcal{L}}} \mathrm{E}_M[\psi] \leqslant \alpha + o(1)$$

as $n \to \infty$, there exists a sequence $\{M_n\}_{n\geqslant 1}$ of models belonging to the class $\mathcal{M}_{\mathcal{L}}$ and a constant $c > 0$ such that for all $n \geqslant 1$, $\inf_{x\in\mathcal{X}} f'_{M_n}(x) < -c(\log n/n)^{\beta/(2\beta+3)}$ and

$$\mathrm{E}_{M_n}[\psi] \leqslant \alpha + o(1)$$

as $n \to \infty$.

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

4. AN ADAPTIVE AND RATE-OPTIMAL TEST

In this section, I study properties of the test based on the basic set of weighting functions and the local Rice estimator. In particular, I show that this test controls asymptotic size and is adaptive and rate optimal against alternatives with regression functions having Lipschitz-continuous derivative. For these purposes, I apply results from the previous section. Specifically, I show that for the test based on the basic set of weighting functions and the local Rice estimator, Assumptions H1–H4 hold uniformly over models $M \in \mathcal{M}_{\mathcal{L}}$. This allows me to apply Theorems 3.1–3.4 with $\mathcal{M}_{\mathcal{H}} = \mathcal{M}_{\mathcal{L}}$ and obtain the desired results.

Throughout this section, I will assume, without further notice, that the basic set of weighting functions is used with a kernel function $K(\cdot)$ having the support $[-1, +1]$, being continuous, and being strictly positive on the interior of its support. Many commonly used kernel functions including uniform, triangular, Epanechnikov, biweight, triweight, and tricube kernels satisfy these restrictions; see Tsybakov (2009) for definitions. Note, however, that these restrictions exclude higher order kernels since those are necessarily negative on parts of their supports.

The following theorem verifies Assumption H1 for the local Rice estimator.

THEOREM 4.1 (Verification of Assumption H1 for local Rice estimator). *Suppose that for all $i = 1, \ldots, n$, $\widehat{\sigma}_i$ is the local Rice estimator of σ_i given in (13) and based on the bandwidth value $b_n = C_b(\log n)^{1/2}/n^{1/4}$ where $C_b > 0$ is some constant. Then Assumption H1 holds uniformly over $M \in \mathcal{M}_{\mathcal{L}}$ for any $\kappa \in (0, 1/4)$.*

Comment 4.1. (*i*) This theorem provides a partial answer to the question of selecting the bandwidth value b_n for the local Rice estimator by showing how fast b_n should converge to zero but it does not provide the full answer because it does not specify the constant C_b in the formula $b_n = C_b(\log n)^{1/2}/n^{1/4}$. In general, the optimal choice of C_b depends on the constants c_1, C_1, c_2, C_2, c_3, C_3, β, and L, which are unknown. Instead, I suggest an ad hoc rule $C_b = 0.2$. With this choice of C_b, the local Rice estimator $\widehat{\sigma}_i$ is based on about 25 observations for all $i = 1, \ldots, n$ when the sample size $n = 100$ and X is distributed uniformly over some interval.
(*ii*) The proof of the theorem also shows that the local Rice estimators $\widehat{\sigma}_i$ satisfy

$$\max_{1 \leqslant i \leqslant n} |\widehat{\sigma}_i - \sigma_i| = O_p\left(b_n + \frac{\log n}{b_n n^{1/2}}\right).$$

This result on the rate of uniform consistency of the local Rice estimators can be of independent interest. □

The following theorem verifies Assumptions H2, H3, and H4 for the basic set of weighting functions.

THEOREM 4.2 (Verification of Assumptions H2, H3, and H4 for basic set of weighting functions). *Suppose that $\mathcal{S}_n$ is the basic set of weighting functions and*

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

that the estimators $\widehat{\sigma}_i$ are the same as those in Theorem 4.1. Then Assumptions H2, H3, and H4 hold with any $\kappa \in (0, 1/4)$ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$.

To conclude this section, I combine Theorems 3.1–3.4 with Theorems 4.1 and 4.2 to obtain the following corollary on the properties of the test based on the basic set of weighting functions and the local Rice estimator.

COROLLARY 4.1 (Size and power properties of test based on the basic set of weighting functions and local Rice estimator). *Let $P = PI$, OS, or SD and suppose that $\mathcal{S}_n$ is the basic set of weighting functions and that the estimators $\widehat{\sigma}_i$ are the same as those in Theorem 4.1. Then*

$$\inf_{M \in \mathcal{M}_{0,\mathcal{L}}} \mathrm{P}_M(T \leqslant c^P_{1-\alpha}) \geqslant 1 - \alpha + o(1) \tag{18}$$

as $n \to \infty$. In addition, let $\mathcal{M}_{00,\mathcal{L}}$ denote the set of all models $M \in \mathcal{M}_{0,\mathcal{L}}$ such that $f_M(\cdot) \equiv C$ for some constant C. Then

$$\sup_{M \in \mathcal{M}_{00,\mathcal{L}}} \left| \mathrm{P}_M(T \leqslant c^P_{1-\alpha}) - (1-\alpha) \right| \to 0 \tag{19}$$

as $n \to \infty$. Furthermore, for any model $M \in \mathcal{M}_{\mathcal{L}}$ such that there exists $x_1, x_2 \in \mathcal{X}$ with $x_1 < x_2$ but $f_M(x_1) > f_M(x_2)$ ($\mathcal{H}_0$ is false),

$$\mathrm{P}_M(T \leqslant c^P_{1-\alpha}) \to 0 \tag{20}$$

as $n \to \infty$. Moreover, if $M \in \mathcal{M}_{\mathcal{L}}$ is such that there exists $x_1, x_2 \in \mathcal{X}$ with $x_1 < x_2$ but $f_M(x_1) > f_M(x_2)$ and $\{M_n\}_{n \geqslant 1}$ is a sequence of models such that for all $n \geqslant 1$, the joint distribution of the pair (X, ε) in M_n coincides with that in M but the regression function $f_{M_n}(\cdot)$ in M_n has the form $f_{M_n}(\cdot) = l_n f_M(\cdot)$, where $\{l_n\}_{n \geqslant 1}$ is some sequence of positive numbers converging to zero ($\mathcal{H}_0$ is false), then

$$\mathrm{P}_{M_n}(T \leqslant c^P_{1-\alpha}) \to 0 \tag{21}$$

as $n \to \infty$ as long as $(\log n/n)^{1/2} = o(l_n)$. Finally, let $\mathcal{M}_{\mathcal{L},l_n}$ denote the set of all models $M \in \mathcal{M}_{\mathcal{L}}$ such that $\inf_{x \in \mathcal{X}} f'_M(x) < -l_n$, where $\{l_n\}_{n \geqslant 1}$ is a sequence of positive numbers converging to zero. Then

$$\sup_{M \in \mathcal{M}_{\mathcal{L},l_n}} \mathrm{P}_M(T \leqslant c^P_{1-\alpha}) \to 0 \tag{22}$$

as $n \to \infty$ as long as $(\log n/n)^{\beta/(2\beta+3)} = o(l_n)$.

Comment 4.2. (*i*) This theorem shows that the test based on the basic set of weighting functions and the local Rice estimator controls asymptotic size, is not conservative, and is consistent against fixed alternatives. The theorem also shows that the test is consistent against local one-dimensional alternatives M_n having the

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

regression function $f_{M_n}(\cdot) = l_n f_M(\cdot)$ for some sequence $\{l_n\}_{n\geqslant 1}$ of positive numbers converging to zero and satisfying $(\log n/n)^{1/2} = o(l_n)$. Finally, the theorem shows that the test is uniformly consistent against alternatives in the set $\mathcal{M}_{\mathcal{L},l_n}$ as long as $(\log n/n)^{\beta/(2\beta+3)} = o(l_n)$.

(*ii*) Comparing the rate of uniform consistency in this theorem against alternatives in the class $\mathcal{M}_{\mathcal{L}}$ with the lower bound on possible rates of uniform consistency in the same class $\mathcal{M}_{\mathcal{L}}$ derived in Theorem 3.5 shows that the optimal rate of uniform consistency is $(\log n/n)^{\beta/(2\beta+3)}$ and that the test based on the basic set of weighting functions and the local Rice estimator is rate optimal. Moreover, carrying out the test does not require knowing β, and so the test is adaptive. Hence, this test is adaptive and rate optimal. □

5. MONTE CARLO SIMULATIONS

In this section, I provide results of a small Monte Carlo simulation study. The purposes of the simulation study are (*i*) to shed some light on the size properties of the tests developed in this article in finite samples, (*ii*) to quantify the power gains from using the one-step and the step-down critical values relative to the plug-in critical values, and (*iii*) to compare the power of the tests developed in this article with that of other tests in the literature. In particular, I consider the tests of Gijbels et al. (2000) (GHJK) and of Ghosal et al. (2000) (GSV).

I consider samples of size $n = 500$, $1,000$, and $2,000$ with X having uniform distribution on the $[-1, 1]$ interval. I also consider the regression functions $f(\cdot)$ of the form

$$f(x) = c_1 x - c_2 \phi(c_3 x), \quad x \in [-1, 1],$$

where c_1, c_2, and c_3 are some parameters, and $\phi(\cdot)$ is the pdf of the standard normal distribution. I assume that ε is a zero-mean random variable that is independent of X. Depending on the experiment, ε has either a Gaussian or a uniform distribution. In both cases, the variance of ε is normalized to be one. Five combinations of the parameter values are studied:

DGP 1: $c_1 = 0, c_2 = 0, c_3 = 0$;
DGP 2: $c_1 = -0.2, c_2 = 0, c_3 = 0$;
DGP 3: $c_1 = 10, c_2 = 50, c_3 = 1$;
DGP 4: $c_1 = 10, c_2 = 10, c_3 = 6$;
DGP 5: $c_1 = 10, c_2 = 8, c_3 = 9$.

DGP 1 satisfies the null hypothesis but DGPs 2–5 do not. DGP 1 has a flat regression function $f(\cdot)$, which gives the least favorable model under the null. DGP 2 has a strictly decreasing regression function $f(\cdot)$, and DGPs 3–5 have regression functions $f(\cdot)$ that are mostly increasing but are strictly decreasing on some parts of the domain. In particular, $f(\cdot)$ is strictly decreasing on $(-1, -0,6)$ in DGP 3, on $(-0.27, -0.08)$ in DGP 4, and on $(-0.19, -0.04)$ in DGP 5. For DGPs 2–5

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

the parameter values were chosen so as to have nontrivial rejection probabilities in most cases (that is, bounded below from zero and above from one).

Let me now describe the tuning parameters for all the tests used in the simulations. For the GHJK and GSV tests, I followed instructions in the corresponding articles. In particular, for the GHJK test, I use their run statistic with $k = 0.2n$ (see the original article for the explanation of the notation), and I use standard normal random variables with $B = 1,000$ bootstrap repetitions to simulate the critical value. For the GSV test, I use their sup-statistic and the kernel function

$$K(x) = 0.75 \cdot (1 - x^2) \cdot 1\{|x| \leqslant 1\}, \quad x \in \mathbb{R} \tag{23}$$

with the bandwidth value $h_n = n^{-1/5}$. To prevent the boundary effects, I take the supremum in the definition of their test statistic over $x \in (-0.9, +0.9)$. Finally, for the tests developed in this article, I use the basic set of weighting functions with the kernel function $K(\cdot)$ given in (23) and the parameter values chosen according to recommendations in Section 2.2 (and $k = 0$). For estimating σ_i's, I use the local Rice estimator with the bandwidth value b_n chosen according to the recommendations in Comment 4.1. For the one-step and the step-down critical values, I take the threshold probability $\gamma_n = 0.01$. All three critical values (plug-in, one-step, and step-down) are calculated using $B = 1,000$ bootstrap repetitions.

For all the tests, I set the nominal level $\alpha = 0.1$. In addition, since my tests slightly over-reject and the GSV test slightly under-reject in finite samples under the null, I use size correction for DGPs 2–5, where the rejection probabilities give the power of the tests. In particular, for these DGPs, I present the results with $\alpha = 0.9$ when ε is Gaussian and $\alpha = 0.8$ when ε is uniform for my tests, and I present the results with $\alpha = 0.13$ both in the Gaussian and in the uniform cases for the GSV test. This correction slightly decreases the power of my tests but slightly increases the power of the GSV test. For each design of the experiment, I use $1,000$ simulations.

Intuitively, the power of the GSV test may be large when the bandwidth value h_n for this test turns out to be appropriate for the given alternative, which happens when the length of the region where the function $f(\cdot)$ is decreasing is of the same order as that of the bandwidth value h_n. In these cases, the power of the GSV test may exceed that of my tests because the former does not incur the cost of adaptation. My tests, however, may have much larger power when the bandwidth value h_n for the GSV test turns out to be inappropriate for the given alternative. Also, as noted in the original article, the advantage of the GHJK test is not its power, which may be low relative to that of other tests, but computational simplicity, minimal requirements on the distribution of ε, and exceptional size control.

The results of the Monte Carlo simulations are presented in Table 1 for the case of the Gaussian ε and in Table 2 for the case of the uniform ε. Specifically, the tables contain information on rejection probabilities for each test depending on the sample size and the DGP. The tests developed in this article are denoted by "ARO, PI", "ARO, OS", and "ARO, SD", where "ARO" is a shorthand for "adaptive and rate optimal". The following observations can be taken from these

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

TABLE 1. Results of monte carlo experiments, Gaussian noise

DGP	Sample Size	Proportion of rejections for				
		GHJK	GSV	ARO, PI	ARO, OS	ARO, SD
1	500	0.115	0.070	0.129	0.129	0.129
	1,000	0.101	0.075	0.130	0.130	0.130
	2,000	0.100	0.074	0.116	0.116	0.116
2	500	0.165	0.154	0.398	0.398	0.398
	1,000	0.148	0.173	0.680	0.680	0.680
	2,000	0.142	0.217	0.913	0.913	0.913
3	500	0.089	0.253	0.220	0.286	0.289
	1,000	0.111	0.456	0.480	0.602	0.605
	2,000	0.143	0.720	0.791	0.830	0.830
4	500	0.081	0.006	0.156	0.226	0.229
	1,000	0.109	0.166	0.371	0.478	0.483
	2,000	0.156	0.767	0.639	0.708	0.709
5	500	0.104	0.000	0.265	0.339	0.340
	1,000	0.162	0.003	0.596	0.685	0.691
	2,000	0.252	0.348	0.904	0.928	0.929

tables. First, results for the Gaussian and the uniform cases are similar. Second, the tests developed in this article may slightly over-reject in moderate samples under the null. For example, in the case of the Gaussian ε, all three tests developed in this article reject with probability 0.129 when $n = 500$ and with probability 0.130 when $n = 1,000$. On the other hand, over-rejection decreases as the sample size grows. In particular, all three tests developed in this article reject with probability 0.116 in the case of the Gaussian ε and with probability 0.106 in the case of the uniform ε when $n = 2,000$. In comparison, the GSV test under-reject under the null, and this under-rejection does not seem to decrease much as the sample size grows. On the other hand, the GHJK test has a very good size control both for the Gaussian and for the uniform ε for all sample sizes considered (in fact, by construction, the GHJK test would have the exact size control if I were to use an infinite number of bootstrap repetitions and an infinite number of simulations). Third, the selection procedure used in the one-step critical value for the tests developed in this article gives improvements in terms of power relative to the plug-in critical value. For example, for DGP 3, the Gaussian ε, and the sample size $n = 1,000$, the rejection probabilities for the "ARO, PI" and the "ARO, OS" tests are 0.480 and 0.602, respectively. On the other hand, the step-down critical value improves power of my tests relative to the one-step critical value only marginally. Fourth, the selection procedures used in the one-step and the step-down critical values for the tests developed in this article do not undermine the size control relative to the plug-in critical value. In particular, for DGP 1, the "ARO, PI", "ARO, OS", and "ARO, SD" tests always have the same rejection probabilities. Fifth, the power of my tests exceeds that of the GHJK and GSV tests in most experiments, sometimes substantially. For example, for DGP 2, the uniform ε, and the sample size $n = 2,000$, all my tests reject with probability 0.907 whereas the GHJK and

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

TABLE 2. Results of monte carlo experiments, uniform noise

DGP	Sample size	Proportion of rejections for				
		GHJK	GSV	ARO, PI	ARO, OS	ARO, SD
1	500	0.117	0.073	0.122	0.122	0.122
	1,000	0.108	0.072	0.111	0.111	0.111
	2,000	0.100	0.076	0.106	0.106	0.106
2	500	0.150	0.150	0.372	0.372	0.372
	1,000	0.146	0.181	0.662	0.662	0.662
	2,000	0.138	0.197	0.907	0.907	0.907
3	500	0.077	0.238	0.204	0.277	0.279
	1,000	0.102	0.459	0.464	0.595	0.596
	2,000	0.138	0.714	0.773	0.812	0.813
4	500	0.075	0.009	0.142	0.204	0.206
	1,000	0.098	0.159	0.352	0.448	0.453
	2,000	0.144	0.732	0.620	0.694	0.696
5	500	0.095	0.000	0.235	0.304	0.306
	1,000	0.134	0.003	0.583	0.674	0.680
	2,000	0.216	0.313	0.899	0.930	0.930

GSV tests reject only with probabilities 0.138 and 0.197, respectively. Finally, the power of the GSV test exceeds that of all my tests only for DGP 4 and the sample size $n = 2,000$, and also exceeds that of my "ARO, PI" test for DGP 3 and the sample size $n = 500$. These are the cases where the bandwidth choice for the GSV test turns out to be appropriate. Even in these cases, however, the difference in the power between the GSV test and my tests is not large. For example, for DGP 4, the uniform ε, and the sample size $n = 2,000$, the GSV test rejects with probability 0.732 and my "ARO, SD" test rejects with probability 0.696.

To complement simulations above, I also compare the rejection probabilities of the adaptive and rate optimal test with step-down critical values with those of the GSV test in the model where X has uniform distribution on the $[-1, 1]$ interval, ε is independent of X and has uniform distribution with mean zero and variance one, and the regression function $f(\cdot)$ takes the following form:

$$f(x) = c_4 x - (10/c_5)\phi(c_5 x), \quad x \in [-1, 1],$$

where c_4 varies between 0 and 3 and c_5 varies between 1 and 10. For both tests, I use the same specifications as those described above. The results are presented in Figure 2, which shows the difference of the rejection probability of the adaptive and rate optimal test and that of the GSV test as a function of c_4 and c_5. To understand the results, note that large values of c_4 correspond to the null hypothesis, and this is the region where the tests have similar rejection probabilities, so that the difference is close to zero. Small values of c_4, on the other hand, correspond to the alternative, and this is the region where the tests have different rejection probabilities. The figure shows that the difference is typically positive and is sometimes substantial in this region, which indicates that the adaptive and rate optimal test developed in this article outperforms the GSV test.

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

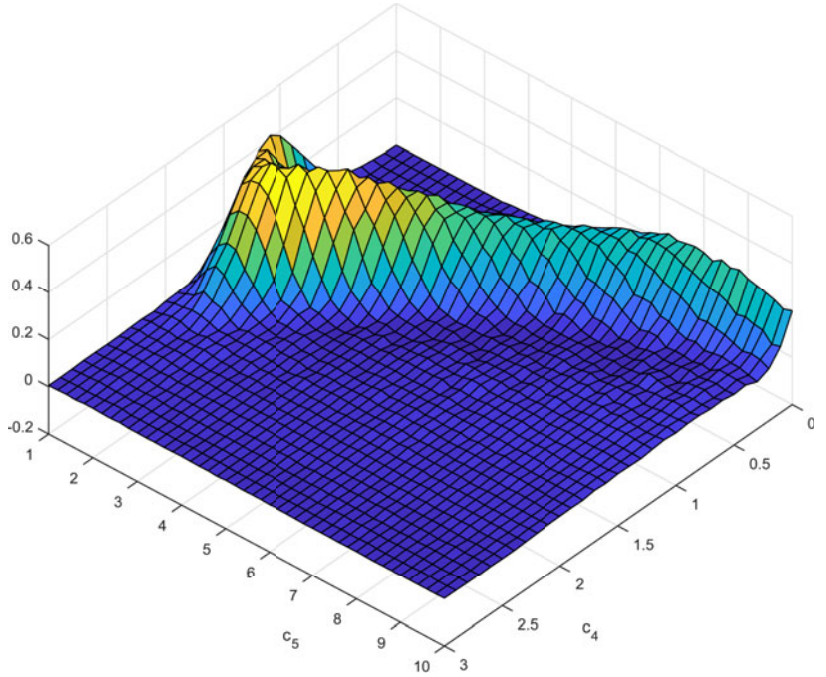

FIGURE 2. Power difference between the adaptive and rate optimal test (with step-down critical values) and the test of Ghosal et al. (2000).

6. CONCLUSION

In this article, I have developed a general framework for testing monotonicity of a nonparametric regression function, and have given a broad class of new tests. A general test statistic uses many different weighting functions so that an approximately optimal weighting function is determined automatically. In this sense, the test adapts to the properties of the model. I have also obtained new methods to simulate the critical values for these tests. These are based on the wild bootstrap and the selection procedures. The selection procedures are used to estimate what counterparts of the test statistic should be used in simulating the critical value. They are constructed so that no violation of the asymptotic size occurs.

REFERENCES

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

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

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

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

Baraud, Y., S. Huet, & B. Laurent (2005) Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *The Annals of Statistics* 33, 214–257.

Bowman, A.W., M.C. Jones, & I. Gijbels (1998) Testing monotonicity of regression. *Journal of Computational and Graphical Statistics* 7, 489–500.

Cai, T. & L. Wang (2008) Adaptive variance function estimation in heteroscedastic nonparametric regression. *The Annals of Statistics* 36, 2025–2054.

Chernozhukov, V., D. Chetverikov, & K. Kato (2013) 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 (2015) 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 (2016a) Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings. *Stochastic Processes and their Applications* 126, 3632–3651.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017) Central limit theorems and bootstrap in high dimensions. The Annals of Probability 45, 2309–2352.

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

Chetverikov, D. (2016) Adaptive test of conditional moment inequalities. *Econometric Theory* 34, 186–227.

Delgado, M. & J. Escanciano (2010) Distribution-free tests of stochastic monotonicity. *Journal of Econometrics* 170, 68–75.

Dudley, R. (1999) *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

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

Durot, C. (2003) A Kolmogorov-type test for monotonicity of regression. *Statistics and Probability Letters* 63, 425–433.

Ellison, G. & S. Ellison (2011) Strategic entry deterrence and the behavior of pharmaceutical incumbents prior to patent expiration. *American Economic Journal: Microeconomics* 3, 1–36.

Fan, J. & Q. Yao (1998) Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* 85, 645–660.

Ghosal, S., A. Sen, & A. van der Vaart (2000) Testing monotonicity of regression. *The Annals of Statistics* 28, 1054–1082.

Gijbels, I., P. Hall, M. Jones, & I. Koch (2000) Tests for monotonicity of a regression mean with guaranteed level. *Biometrika* 87, 663–673.

Gutknecht (2016) Testing for monotonicity under endogeneity - An application to the reservation wage function. *Journal of Econometrics* 190, 100–114.

Hall, P. & N. Heckman (2000) Testing for monotonicity of a regression mean by calibrating for linear functions. *The Annals of Statistics* 28, 20–39.

Hardle, W. & E. Mammen (1993) Comparing nonparametric versus parametric regression fits. *The Annals of Statistics* 21, 1926–1947.

Hardle, W. & A. Tsybakov (2007) Local polinomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics* 81, 233–242.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65–70.

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.

Juditsky, A. & A. Nemirovski (2002) On nonparametric tests of positivity/monotonicity/convexity. *The Annals of Statistics* 30, 498–527.
Lehmann, E.L. & J. Romano (2005) *Testing Statistical Hypotheses*. Springer.
Lee, S., O. Linton, & Y. Whang (2009) Testing for stochastic monotonicity. *Econometrica* 27, 585–602.
Lee, S., K. Song, and Y.-J. Whang (2017) Testing for a general class of functional inequalities. *Econometric Theory*, 1–47.
Liu, R. (1988) Bootstrap procedures under niid models. *The Annals of Statistics* 16, 1696–1708.
Mammen, E. (1993) Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics* 21, 255–285.
Matzkin, R. (1994) Restrictions of economic theory in nonparametric methods. *Handbook of Econometrics, Volume IV*. Edited by R. Engle and D. McFadden, Elsevier Science, 2523–2558.
Milgrom, P. & C. Shannon (1994) Monotone comparative statics. *Econometrica* 62, 157–180.
Muller, H. & U. Stadtmuller (1987) Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics* 15, 610–625.
Rice, J. (1984) Bandwidth choice for nonparametric kernel regression. *The Annals of Statistics* 12, 1215–1230.
Romano, J. & A. Shaikh (2010) Inference for the identified sets in partially identified econometric models. *Econometrica* 78, 169–211.
Romano, J. & M. Wolf (2005a) Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of American Statistical Association* 100, 94–108.
Romano, J. & M. Wolf (2005b) Stepwise multiple testing as formalized data snooping. *Econometrica* 73, 1237–1282.
Romano, J. & M. Wolf (2013) Testing for monotonicity in expected asset returns. *Journal of Empirical Finance* 23, 93–116.
Schlee (1982) Nonparametric tests of the monotonicity and convexity of regression. *Nonparametric Statistical Inference, Volume II*. Edited by B. Gnedenko, M. Puri, and I. Vincze, North-Holland, 823–836.
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.
Wang, J. & M. Meyer (2011) Testing the monotonicity or convexity of a function using regression splines. *The Canadian Journal of Statistics* 39, 89–107.
Wu, C. (1986) Jacknife, bootstrap, and other resampling methods in regression analysis. *The Annals of Statistics* 14, 1261–1295.

APPENDIX A: Proofs for Section 3

In this appendix, I first prove a sequence of auxiliary lemmas (Section A.1) and then I present the proofs of the theorems stated in Section 3 (Section A.2). In Section A.1, all results hold uniformly over models $M \in \mathcal{M}_{\mathcal{H}}$. For brevity, however, I drop the index M, and do not claim this uniformity repeatedly.

Throughout this appendix, I will use the following additional notation. Let

$$w_i(s) = \sum_{1 \leqslant j \leqslant n} \text{sign}(X_j - X_i) Q(X_i, X_j, s), \quad i = 1, \ldots, n, \ s \in \mathcal{S}_n,$$

$$e(s) = \sum_{1 \leqslant i \leqslant n} \frac{w_i(s)\sigma_i \epsilon_i}{(V(s))^{1/2}}, \text{ and } \varepsilon(s) = \sum_{1 \leqslant i \leqslant n} \frac{w_i(s)\varepsilon_i}{(V(s))^{1/2}}, \quad s \in \mathcal{S}_n,$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

where $\{\epsilon_i\}_{1\leqslant i\leqslant n}$ are i.i.d. $N(0,1)$ random variables that are independent of the data $\{X_i, Y_i\}_{1\leqslant i\leqslant n}$. Also, let

$$\mathcal{V}_n = \max_{s\in\mathcal{S}_n} (V(s)/\widehat{V}(s))^{1/2}.$$

Moreover, for all $\mathcal{S}\subset\mathcal{S}_n$ and $\eta\in(0,1)$, let $c_\eta^{\mathcal{S}}$ denote the η quantile of the conditional distribution of

$$T^{\mathcal{S}} = \max_{s\in\mathcal{S}} \sum_{1\leqslant i\leqslant n} \frac{w_i(s)\widehat{\sigma}_i\epsilon_i}{(\widehat{V}(s))^{1/2}}$$

given the data $\{X_i, Y_i\}_{1\leqslant i\leqslant n}$, and let $c_\eta^{\mathcal{S},0}$ denote the η quantile of the conditional distribution of

$$T^{\mathcal{S},0} = \max_{s\in\mathcal{S}} \sum_{1\leqslant i\leqslant n} \frac{w_i(s)\sigma_i\epsilon_i}{(V(s))^{1/2}} = \max_{s\in\mathcal{S}} e(s)$$

given the data $\{X_i, Y_i\}_{1\leqslant i\leqslant n}$. Finally, for $\eta\leqslant 0$, define $c_\eta^{\mathcal{S}}$ and $c_\eta^{\mathcal{S},0}$ as $-\infty$, and for $\eta\geqslant 1$, define $c_\eta^{\mathcal{S}}$ and $c_\eta^{\mathcal{S},0}$ as $+\infty$.

A.1. Auxiliary Lemmas

LEMMA A.1. *For some universal constant $C>0$,*

$$\mathrm{E}\left[\max_{s\in\mathcal{S}_n} |e(s)| \,|\, \{X_i\}_{1\leqslant i\leqslant n}\right] \leqslant C(\log p)^{1/2}.$$

Proof. For all $s\in\mathcal{S}_n$, conditional on $\{X_i\}_{1\leqslant i\leqslant n}$, the random variable $e(s)$ has the $N(0,1)$ distribution, and $|\mathcal{S}_n| = p$. So, the result follows from Lemma 2.2.2 in van der Vaart and Wellner (1996). ■

LEMMA A.2. *Let $\mathcal{S}\subset\mathcal{S}_n$. Then for all $\Delta>0$,*

$$\sup_{t\in\mathbb{R}} \mathrm{P}\left(\max_{s\in\mathcal{S}} e(s)\in[t,t+\Delta] \,|\, \{X_i\}_{1\leqslant i\leqslant n}\right) \leqslant C\Delta(\log p)^{1/2},$$

and for all $(\eta,\delta)\in(0,1)^2$,

$$c_{\eta+\delta}^{\mathcal{S},0} - c_\eta^{\mathcal{S},0} \geqslant \frac{\delta}{C(\log p)^{1/2}} \tag{A.1}$$

for some universal constant $C>0$.

Proof. Theorem 3 in Chernozhukov et al. (2015) shows that if $W_1,\ldots,W_p$ are $N(0,1)$ random variables (not necessarily independent), then for all $\Delta>0$,

$$\sup_{t\in\mathbb{R}} \mathrm{P}\left(\max_{1\leqslant j\leqslant p} W_j \in [t,t+\Delta]\right) \leqslant 4\Delta\left(\mathrm{E}\left[\max_{1\leqslant j\leqslant p} W_j\right] + 1\right).$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

Therefore,

$$\sup_{t\in\mathbb{R}} \mathrm{P}\left(\max_{s\in\mathcal{S}} e(s) \in [t, t+\Delta] | \{X_i\}_{1\leqslant i\leqslant n}\right) \leqslant 4\Delta\left(\mathrm{E}\left[\max_{s\in\mathcal{S}} e(s) | \{X_i\}_{1\leqslant i\leqslant n}\right] + 1\right),$$

and so the first asserted claim follows from Lemma A.1. For the second asserted claim, note that (A.1) holds if $\eta+\delta \geqslant 1$ since $c_{\eta+\delta}^{\mathcal{S},0} = +\infty$ in this case. On the other hand, if $\eta+\delta < 1$, then

$$\delta = \mathrm{P}\left(\max_{s\in\mathcal{S}} e(s) \in [c_{\eta}^{\mathcal{S},0}, c_{\eta+\delta}^{\mathcal{S},0}] | \{X_i\}_{1\leqslant i\leqslant n}\right) \leqslant C(c_{\eta+\delta}^{\mathcal{S},0} - c_{\eta}^{\mathcal{S},0})(\log p)^{1/2},$$

and so (A.1) holds as well. This gives the second asserted claim and completes the proof. ■

LEMMA A.3. *For all $\mathcal{S} \subset \mathcal{S}_n$ and $\eta \in (0,1)$,*

$$c_{\eta}^{\mathcal{S},0} \leqslant \frac{C(\log p)^{1/2}}{1-\eta} \tag{A.2}$$

for some universal constant $C > 0$.

Proof. Recall that $c_{\eta}^{\mathcal{S},0}$ is the η quantile of the conditional distribution of $\max_{s\in\mathcal{S}} e(s)$ given $\{X_i\}_{1\leqslant i\leqslant n}$. Hence, by Markov's inequality and Lemma A.1,

$$1-\eta = \mathrm{P}\left[\max_{s\in\mathcal{S}} e(s) > c_{\eta}^{\mathcal{S},0} | \{X_i\}_{1\leqslant i\leqslant n}\right] \leqslant \frac{\mathrm{E}[\max_{s\in\mathcal{S}} e(s) | \{X_i\}_{1\leqslant i\leqslant n}]}{c_{\eta}^{\mathcal{S},0}} \leqslant \frac{C(\log p)^{1/2}}{c_{\eta}^{\mathcal{S},0}}$$

for some universal constant $C > 0$. Rearranging this inequality gives the asserted claim. ■

LEMMA A.4. *The random variable*

$$U_n = \max_{s\in\mathcal{S}_n} \left| \sum_{1\leqslant i\leqslant n} \left(\frac{w_i(s)\widehat{\sigma}_i\epsilon_i}{(\widehat{V}(s))^{1/2}} - \frac{w_i(s)\sigma_i\epsilon_i}{(V(s))^{1/2}} \right) \right|$$

satisfies

$$U_n = o_p\left((\log p)^{1/2}/n^{\kappa}\right).$$

In addition, there exists a sequence $\{\widetilde{\psi}_n\}_{n\geqslant 1}$ of positive numbers converging to zero such that

$$\mathrm{P}\left(U_n > (\log p)^{1/2}/n^{\kappa}\right) = o(\widetilde{\psi}_n).$$

Proof. Denote

$$U_{1,n} = \max_{s\in\mathcal{S}_n} \left| \sum_{1\leqslant i\leqslant n} \frac{w_i(s)\sigma_i\epsilon_i}{(V(s))^{1/2}} \right| \times \max_{s\in\mathcal{S}_n} \left| 1 - (V(s)/\widehat{V}(s))^{1/2} \right| \text{ and}$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

$$U_{2,n} = \max_{s\in\mathcal{S}_n}\left|\sum_{1\leqslant i\leqslant n}\frac{w_i(s)(\widehat{\sigma}_i-\sigma_i)\epsilon_i}{(V(s))^{1/2}}\right| \times \max_{s\in\mathcal{S}_n}(V(s)/\widehat{V}(s))^{1/2}.$$

Then

$$U_n \leqslant U_{1,n} + U_{2,n}, \tag{A.3}$$

and so it suffices to bound $U_{1,n}$ and $U_{2,n}$ separately. To bound $U_{1,n}$ note that by Lemma A.1,

$$\mathrm{E}\left[\max_{s\in\mathcal{S}_n}\left|\sum_{1\leqslant i\leqslant n}\frac{w_i(s)\sigma_i\epsilon_i}{(V(s))^{1/2}}\right|\right] = \mathrm{E}\left[\max_{s\in\mathcal{S}_n}|e(s)|\right] = \mathrm{E}\left[\mathrm{E}\left[\max_{s\in\mathcal{S}_n}|e(s)||\{X_i\}_{1\leqslant i\leqslant n}\right]\right] \leqslant C(\log p)^{1/2}$$

for some universal constant $C > 0$ and by Assumption H2,

$$\max_{s\in\mathcal{S}_n}\left|1-(V(s)/\widehat{V}(s))^{1/2}\right| = o_p(n^{-\kappa}).$$

Combining these bounds shows that

$$U_{1,n} = o_p\left((\log p)^{1/2}/n^{\kappa}\right). \tag{A.4}$$

Furthermore, to bound $U_{2,n}$, note that by Assumptions L1 and H1, conditional on the data $\{X_i, Y_i\}_{1\leqslant i\leqslant n}$, the random variable

$$\sum_{1\leqslant i\leqslant n}\frac{w_i(s)(\widehat{\sigma}_i-\sigma_i)\epsilon_i}{(V(s))^{1/2}}$$

has a zero-mean Gaussian distribution with variance bounded from above by

$$\sum_{1\leqslant i\leqslant n}\frac{w_i(s)^2(\widehat{\sigma}_i-\sigma_i)^2}{V(s)} \leqslant \frac{\max_{1\leqslant i\leqslant n}|\widehat{\sigma}_i-\sigma_i|^2}{\min_{1\leqslant i\leqslant n}\sigma_i^2} \times \sum_{1\leqslant i\leqslant n}\frac{w_i(s)^2\sigma_i^2}{V(s)}$$
$$= \frac{\max_{1\leqslant i\leqslant n}|\widehat{\sigma}_i-\sigma_i|^2}{\min_{1\leqslant i\leqslant n}\sigma_i^2} = o_p(n^{-2\kappa}).$$

Hence, by the same argument as that used in the proof of Lemma A.1 and Markov's inequality,

$$\max_{s\in\mathcal{S}_n}\left|\sum_{1\leqslant i\leqslant n}\frac{w_i(s)(\widehat{\sigma}_i-\sigma_i)\epsilon_i}{(V(s))^{1/2}}\right| = o_p\left((\log p)^{1/2}/n^{\kappa}\right).$$

Since $\max_{s\in\mathcal{S}}(V(s)/\widehat{V}(s))^{1/2} \to_p 1$ by Assumption H2, the last bound implies that

$$U_{2,n} = o_p\left((\log p)^{1/2}/n^{\kappa}\right). \tag{A.5}$$

Combining (A.3), (A.4), and (A.5) gives the first asserted claim. The second asserted claim follows from the first claim. ■

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

LEMMA A.5. *There exists an event $\mathcal{A}_n$ depending on the data only via $\{X_i\}_{1\leqslant i\leqslant n}$ such that (i) $\mathrm{P}(\mathcal{A}_n) = 1 - o(1)$ and (ii) uniformly over $\{X_i\}_{1\leqslant i\leqslant n} \in \mathcal{A}_n$, $\mathcal{S} \subset \mathcal{S}_n$, and $\eta \in (0,1)$,*

$$\mathrm{P}\left(\max_{s\in\mathcal{S}} \varepsilon(s) \leqslant c_\eta^{\mathcal{S},0} | \{X_i\}_{1\leqslant i\leqslant n}\right) = \eta + o(1)$$

and

$$\mathrm{P}\left(\max_{s\in\mathcal{S}} (-\varepsilon(s)) \leqslant c_\eta^{\mathcal{S},0} | \{X_i\}_{1\leqslant i\leqslant n}\right) = \eta + o(1).$$

Proof. Recall that

$$\varepsilon(s) = \sum_{1\leqslant i\leqslant n} \frac{w_i(s)\varepsilon_i}{(V(s))^{1/2}}, \quad s \in \mathcal{S}_n.$$

Also, note that for all $s \in \mathcal{S}$,

$$\mathrm{E}[\varepsilon(s)^2] = \sum_{1\leqslant i\leqslant n} \frac{(w_i(s)\sigma_i)^2}{V(s)} = 1.$$

In addition, by Assumption H3, $nA_n^4 \log^7(pn) = o_p(1)$. Hence, there exists a sequence $\{\omega_n\}_{n\geqslant 1}$ of positive numbers such that $nA_n^4 \log^7(pn) = o_p(\omega_n)$ and $\omega_n = o(1)$. Let $\mathcal{A}_n$ denote the event that $nA_n^4 \log^7(pn) \leqslant \omega_n$. Then $\mathrm{P}(\mathcal{A}_n) = 1 - o(1)$, and the event $\mathcal{A}_n$ depends on the data only via $\{X_i\}_{1\leqslant i\leqslant n}$. Furthermore, note that $\mathrm{E}[\varepsilon_i^2/\sigma_i^2 | X_i] = 1$, and since $(\mathrm{E}[\varepsilon_i^4 | X_i])^{1/4} \leqslant C_1$ and $\sigma_i \geqslant c_1$ by Assumption L1, $\mathrm{E}[\varepsilon_i^4/\sigma_i^4 | X_i] \leqslant C$ for some constant $C > 0$. Therefore, applying Lemma C.3 with $z_{is} = \sqrt{n} w_i(s)\sigma_i/(V(s))^{1/2}$ and $u_i = \varepsilon_i/\sigma_i$ conditional on $\{X_i\}_{1\leqslant i\leqslant n}$ shows that there exists a sequence $\{\omega'\}_{n\geqslant 1}$ of positive numbers converging to zero and depending only on $\{\omega_n\}_{n\geqslant 1}$ such that on the event $\mathcal{A}_n$,

$$\left|\mathrm{P}\left(\max_{s\in\mathcal{S}} \varepsilon(s) \leqslant c_\eta^{\mathcal{S},0} | \{X_i\}_{1\leqslant i\leqslant n}\right) - \eta\right| \leqslant \omega_n'.$$

The first asserted claim follows. The second asserted claim follows by replacing $\varepsilon(s)$ by $-\varepsilon(s)$. ∎

LEMMA A.6. *There exist a sequence $\{\psi_n\}_{n\geqslant 1}$ of positive numbers converging to zero and an event $\mathcal{B}_n$ depending on the data only via $\{X_i\}_{1\leqslant i\leqslant n}$ such that (i) $\mathrm{P}(\mathcal{B}_n) = 1 - o(1)$ and (ii) uniformly over $\{X_i\}_{1\leqslant i\leqslant n} \in \mathcal{B}_n$, $\mathcal{S} \subset \mathcal{S}_n$, and $\eta \in (0,1)$,*

$$\mathrm{P}(c_{\eta+\psi_n}^{\mathcal{S},0} < c_\eta^{\mathcal{S}} | \{X_i\}_{1\leqslant i\leqslant n}) = o(1) \text{ and } \mathrm{P}(c_{\eta+\psi_n}^{\mathcal{S}} < c_\eta^{\mathcal{S},0} | \{X_i\}_{1\leqslant i\leqslant n}) = o(1).$$

Proof. By Assumption H3, $\log p/n^\kappa = o(1)$. Hence, there exists a sequence $\{\omega_n\}_{n\geqslant 1}$ of positive numbers converging to zero such that $\log p/n^\kappa \leqslant \omega_n$. Furthermore, recall the random variables $T^{\mathcal{S}}$ and $T^{\mathcal{S},0}$ defined in the beginning of this appendix:

$$T^{\mathcal{S}} = \max_{s\in\mathcal{S}} \sum_{1\leqslant i\leqslant n} \frac{w_i(s)\widehat{\sigma}_i\epsilon_i}{(\widehat{V}(s))^{1/2}} \text{ and } T^{\mathcal{S},0} = \max_{s\in\mathcal{S}} \sum_{1\leqslant i\leqslant n} \frac{w_i(s)\sigma_i\epsilon_i}{(V(s))^{1/2}}.$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

Then for all $\mathcal{S} \subset \mathcal{S}_n$,

$$|T^{\mathcal{S}} - T^{\mathcal{S},0}| \leqslant \max_{s \in \mathcal{S}} \left| \sum_{1 \leqslant i \leqslant n} \left(\frac{w_i(s)\widehat{\sigma}_i \epsilon_i}{(\widehat{V}(s))^{1/2}} - \frac{w_i(s)\sigma_i \epsilon_i}{(V(s))^{1/2}} \right) \right|$$
$$\leqslant \max_{s \in \mathcal{S}_n} \left| \sum_{1 \leqslant i \leqslant n} \left(\frac{w_i(s)\widehat{\sigma}_i \epsilon_i}{(\widehat{V}(s))^{1/2}} - \frac{w_i(s)\sigma_i \epsilon_i}{(V(s))^{1/2}} \right) \right| = U_n$$

for U_n defined in the statement of Lemma A.4. Therefore, Lemma A.4 shows that there exists a sequence $\{\tilde{\psi}_n\}$ of positive numbers converging to zero such that

$$\mathrm{P}\left(\max_{\mathcal{S} \subset \mathcal{S}_n} |T^{\mathcal{S}} - T^{\mathcal{S},0}| > (\log p)^{1/2}/n^{\kappa} \right) = o(\tilde{\psi}_n).$$

Hence,

$$\mathrm{P}\left(\mathrm{P}\left(\max_{\mathcal{S} \subset \mathcal{S}_n} |T^{\mathcal{S}} - T^{\mathcal{S},0}| > (\log p)^{1/2}/n^{\kappa} | \{X_i, Y_i\}_{1 \leqslant i \leqslant n} \right) > \tilde{\psi}_n \right) \to 0.$$

Let $\overline{\mathcal{B}}_n$ denote the event that

$$\mathrm{P}\left(\max_{\mathcal{S} \subset \mathcal{S}_n} |T^{\mathcal{S}} - T^{\mathcal{S},0}| > (\log p)^{1/2}/n^{\kappa} | \{X_i, Y_i\}_{1 \leqslant i \leqslant n} \right) \leqslant \tilde{\psi}_n,$$

so that

$$\mathrm{P}(\overline{\mathcal{B}}_n) = 1 - o(1). \tag{A.6}$$

Define $\psi_n = \tilde{\psi}_n + \omega_n^{1/2} + C\omega_n$ where C is the same universal constant as that in (A.1) in the statement of Lemma A.2. Then $\psi_n = o(1)$. Also, observe that

$$\mathrm{P}\left(T^{\mathcal{S},0} \leqslant c_{\eta}^{\mathcal{S},0} | \{X_i, Y_i\}_{1 \leqslant i \leqslant n} \right) = \eta \text{ and } \mathrm{P}\left(T^{\mathcal{S}} \leqslant c_{\eta}^{\mathcal{S}} | \{X_i, Y_i\}_{1 \leqslant i \leqslant n} \right) = \eta$$

for any $\eta \in (0,1)$. So, on $\overline{\mathcal{B}}_n$, for all $\mathcal{S} \subset \mathcal{S}_n$ and $\eta \in (0,1)$ such that $\eta + \tilde{\psi}_n < 1$,

$$\eta + \tilde{\psi}_n = \mathrm{P}\left(T^{\mathcal{S},0} \leqslant c_{\eta+\tilde{\psi}_n}^{\mathcal{S},0} | \{X_i, Y_i\}_{1 \leqslant i \leqslant n} \right)$$
$$\leqslant \mathrm{P}\left(T^{\mathcal{S}} \leqslant c_{\eta+\tilde{\psi}_n}^{\mathcal{S},0} + \omega_n/(\log p)^{1/2} | \{X_i, Y_i\}_{1 \leqslant i \leqslant n} \right) + \tilde{\psi}_n$$
$$\leqslant \mathrm{P}\left(T^{\mathcal{S}} \leqslant c_{\eta+\psi_n}^{\mathcal{S},0} | \{X_i, Y_i\}_{1 \leqslant i \leqslant n} \right) + \tilde{\psi}_n,$$

where the last inequality follows from Lemma A.2. Therefore, $c_{\eta}^{\mathcal{S}} \leqslant c_{\eta+\psi_n}^{\mathcal{S},0}$ for all $\mathcal{S} \subset \mathcal{S}_n$ and $\eta \in (0,1)$ such that $\eta + \tilde{\psi}_n < 1$ on the event $\overline{\mathcal{B}}_n$. On the other hand, if $\eta + \tilde{\psi}_n \geqslant 1$, then $\eta + \psi_n \geqslant 1$ and $c_{\eta}^{\mathcal{S}} \leqslant c_{\eta+\psi_n}^{\mathcal{S},0}$ for all $\mathcal{S} \subset \mathcal{S}_n$ and $\eta \in (0,1)$. Conclude that $c_{\eta}^{\mathcal{S}} \leqslant c_{\eta+\psi_n}^{\mathcal{S},0}$ for all $\mathcal{S} \subset \mathcal{S}_n$ and $\eta \in (0,1)$ on the event $\overline{\mathcal{B}}_n$ (both when $\eta + \tilde{\psi}_n < 1$ and when $\eta + \tilde{\psi}_n \geqslant 1$). In addition, by (A.6), it follows that there exists an event $\mathcal{B}_n$ depending on the data only via $\{X_i\}_{1 \leqslant i \leqslant n}$ such that (i) $\mathrm{P}(\mathcal{B}_n) = 1 - o(1)$ and (ii) uniformly over $\{X_i\}_{1 \leqslant i \leqslant n} \in \mathcal{B}_n$, $\mathrm{P}(\overline{\mathcal{B}}_n | \{X_i\}) = 1 - o(1)$. The first asserted claim follows. The second asserted claim follows similarly. ■

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

LEMMA A.7. *Define*

$$\mathcal{S}_n^R = \left\{ s \in \mathcal{S}_n : \sum_{1 \leqslant i \leqslant n} \frac{w_i(s) f(X_i)}{(V(s))^{1/2}} > -c_{1-\gamma_n-\psi_n}^{\mathcal{S}_n,0} \right\},$$

where $\{\psi_n\}_{n \geqslant 1}$ *is the sequence of positive numbers defined in the statement of Lemma A.6. Then* $\mathrm{P}(\mathcal{S}_n^R \subset \mathcal{S}_n^{SD}) \geqslant 1 - \gamma_n + o(1)$ *and* $\mathrm{P}(\mathcal{S}_n^R \subset \mathcal{S}_n^{OS}) \geqslant 1 - \gamma_n + o(1)$.

Proof. I start with some preliminary bounds. First, by Lemma A.6,

$$\mathrm{P}(c_{1-\gamma_n-\psi_n}^{\mathcal{S}_n,0} > c_{1-\gamma_n}^{\mathcal{S}_n}) = o(1) \text{ and } \mathrm{P}(c_{1-\gamma_n-\psi_n}^{\mathcal{S}_n^R,0} > c_{1-\gamma_n}^{\mathcal{S}_n^R}) = o(1). \tag{A.7}$$

Second, from the proof of Lemma A.6, $\log p / n^{\kappa} = o(\psi_n)$. Hence, there exists a sequence $\{\omega_n\}_{n \geqslant 1}$ of positive numbers converging to zero such that

$$\log p / n^{\kappa} \leqslant \omega_n \psi_n. \tag{A.8}$$

Third, recall the random variable $\mathcal{V}_n$ defined in the beginning of this appendix. By Lemma A.3 and Assumption H2,

$$\mathrm{P}\left(\left| c_{1-\gamma_n-\psi_n}^{\mathcal{S}_n,0} (1/\mathcal{V}_n - 1) \right| > \frac{(\log p)^{1/2}}{n^{\kappa}(\gamma_n + \psi_n)} \right) = o(1). \tag{A.9}$$

Fourth, by Lemma A.5, uniformly over $\eta \in (0,1)$,

$$\mathrm{P}\left(\max_{s \in \mathcal{S}_n^R} (-\varepsilon(s)) \geqslant c_{1-\eta}^{\mathcal{S}_n^R,0} \right) = \eta + o(1). \tag{A.10}$$

Finally, suppose that $\mathcal{S}_n^R \backslash \mathcal{S}_n^{SD} \neq \emptyset$. Then there exists the smallest integer l such that $\mathcal{S}_n^R \backslash \mathcal{S}_n^{l-1} = \emptyset$ but $\mathcal{S}_n^R \backslash \mathcal{S}_n^l \neq \emptyset$ (if $l = 1$, let $\mathcal{S}_n^0 = \mathcal{S}_n$). For this l, $c_{1-\gamma_n}^{\mathcal{S}_n^R} \leqslant c_{1-\gamma_n}^{\mathcal{S}_n^{l-1}} = c_{1-\gamma_n}^{l-1}$. Therefore, there exists $s \in \mathcal{S}_n^R$ such that

$$\sum_{1 \leqslant i \leqslant n} \frac{w_i(s) Y_i}{(\widehat{V}(s))^{1/2}} \leqslant -c_{1-\gamma_n}^{PI} - c_{1-\gamma_n}^{l-1} \leqslant -c_{1-\gamma_n}^{PI} - c_{1-\gamma_n}^{\mathcal{S}_n^R} = -c_{1-\gamma_n}^{\mathcal{S}_n} - c_{1-\gamma_n}^{\mathcal{S}_n^R}.$$

It follows that

$$\begin{aligned}
\mathrm{P}(\mathcal{S}_n^R \backslash \mathcal{S}_n^{SD} \neq \emptyset) &\leqslant \mathrm{P}\Big(\min_{s \in \mathcal{S}_n^R} \textstyle\sum_{1 \leqslant i \leqslant n} w_i(s) Y_i / (\widehat{V}(s))^{1/2} \leqslant -c_{1-\gamma_n}^{\mathcal{S}_n} - c_{1-\gamma_n}^{\mathcal{S}_n^R} \Big) \\
&\leqslant_{(1)} \mathrm{P}\Big(\mathcal{V}_n \min_{s \in \mathcal{S}_n^R} \textstyle\sum_{1 \leqslant i \leqslant n} w_i(s) Y_i / (V(s))^{1/2} \leqslant -c_{1-\gamma_n}^{\mathcal{S}_n} - c_{1-\gamma_n}^{\mathcal{S}_n^R} \Big) + o(1) \\
&\leqslant_{(2)} \mathrm{P}\Big(\mathcal{V}_n \min_{s \in \mathcal{S}_n^R} \textstyle\sum_{1 \leqslant i \leqslant n} w_i(s) Y_i / (V(s))^{1/2} \leqslant -c_{1-\gamma_n-\psi_n}^{\mathcal{S}_n,0} - c_{1-\gamma_n-\psi_n}^{\mathcal{S}_n^R,0} \Big) + o(1) \\
&\leqslant_{(3)} \mathrm{P}\Big(\mathcal{V}_n \min_{s \in \mathcal{S}_n^R} (\varepsilon(s) - c_{1-\gamma_n-\psi_n}^{\mathcal{S}_n,0}) \leqslant -c_{1-\gamma_n-\psi_n}^{\mathcal{S}_n,0} - c_{1-\gamma_n-\psi_n}^{\mathcal{S}_n^R,0} \Big) + o(1),
\end{aligned}$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

where *(1)* is by the definition of $\mathcal{V}_n$ since $c^{\mathcal{S}_n}_{1-\gamma_n}+c^{\mathcal{S}^R_n}_{1-\gamma_n}\geqslant 0$ for n large enough (remember that $\gamma_n\to 0$ as $n\to\infty$), *(2)* is by (A.7), and *(3)* is by the definition of $\mathcal{S}^R_n$. Furthermore, for some constant $C'>0$, the right-hand side of *(3)* is equal to

$$
\begin{aligned}
&\mathrm{P}\Big(\max_{s\in\mathcal{S}^R_n}(-\varepsilon(s))\geqslant c^{\mathcal{S}_n,0}_{1-\gamma_n-\psi_n}(1/\mathcal{V}_n-1)+c^{\mathcal{S}^R_n,0}_{1-\gamma_n-\psi_n}/\mathcal{V}_n\Big)+o(1)\\
&\quad\leqslant_{(4)}\mathrm{P}\Big(\max_{s\in\mathcal{S}^R_n}(-\varepsilon(s))\geqslant c^{\mathcal{S}^R_n,0}_{1-\gamma_n-\psi_n}/\mathcal{V}_n-(\log p)^{1/2}n^{-\kappa}/(\gamma_n+\psi_n)\Big)+o(1)\\
&\quad\leqslant_{(5)}\mathrm{P}\Big(\max_{s\in\mathcal{S}^R_n}(-\varepsilon(s))\geqslant c^{\mathcal{S}^R_n,0}_{1-\gamma_n-\psi_n}-2(\log p)^{1/2}n^{-\kappa}/(\gamma_n+\psi_n)\Big)+o(1)\\
&\quad\leqslant_{(6)}\mathrm{P}\Big(\max_{s\in\mathcal{S}^R_n}(-\varepsilon(s))\geqslant c^{\mathcal{S}^R_n,0}_{1-\gamma_n-\psi_n-C'(\log p)n^{-\kappa}/(\gamma_n+\psi_n)}\Big)+o(1)\\
&\quad\leqslant_{(7)}\mathrm{P}\Big(\max_{s\in\mathcal{S}^R_n}(-\varepsilon(s))\geqslant c^{\mathcal{S}^R_n,0}_{1-\gamma_n-\psi_n-C'\omega_n\psi_n/(\gamma_n+\psi_n)}\Big)+o(1)\\
&\quad=_{(8)}\gamma_n+\psi_n+C'\omega_n\psi_n/(\gamma_n+\psi_n)+o(1)=_{(9)}\gamma_n+o(1),
\end{aligned}
$$

where *(4)* is by (A.9), *(5)* is by (A.9) since $c^{\mathcal{S}^R_n,0}_{1-\gamma_n-\psi_n}\leqslant c^{\mathcal{S}_n,0}_{1-\gamma_n-\psi_n}$, *(6)* is by Lemma A.2, *(7)* is by (A.8), *(8)* is by (A.10), and *(9)* is by the definitions of ψ_n and ω_n. The first asserted claim follows. The second claim follows from the fact that $\mathcal{S}^{SD}_n\subset\mathcal{S}^{OS}_n$. ■

LEMMA A.8. *The set $\mathcal{S}^R_n$ defined in the statement of Lemma A.7 satisfies*

$$
\mathrm{P}\left(\max_{s\in\mathcal{S}_n\setminus\mathcal{S}^R_n}\sum_{1\leqslant i\leqslant n}\frac{w_i(s)Y_i}{(\widehat{V}(s))^{1/2}}\leqslant 0\right)\geqslant 1-\gamma_n+o(1).
$$

Proof. The asserted claim follows from

$$
\begin{aligned}
&\mathrm{P}\Big(\max_{s\in\mathcal{S}_n\setminus\mathcal{S}^R_n}\textstyle\sum_{1\leqslant i\leqslant n}w_i(s)Y_i/(\widehat{V}(s))^{1/2}\leqslant 0\Big)=\mathrm{P}\Big(\max_{s\in\mathcal{S}_n\setminus\mathcal{S}^R_n}\textstyle\sum_{1\leqslant i\leqslant n}w_i(s)Y_i/(V(s))^{1/2}\leqslant 0\Big)\\
&\quad\geqslant_{(1)}\mathrm{P}\Big(\max_{s\in\mathcal{S}_n\setminus\mathcal{S}^R_n}\textstyle\sum_{1\leqslant i\leqslant n}w_i(s)\varepsilon_i/(V(s))^{1/2}\leqslant c^{PI,0}_{1-\gamma_n-\psi_n}\Big)=_{(2)}\mathrm{P}\Big(\max_{s\in\mathcal{S}_n}\varepsilon(s)\leqslant c^{PI,0}_{1-\gamma_n-\psi_n}\Big)\\
&\quad=_{(3)}\mathrm{P}\Big(\max_{s\in\mathcal{S}_n}\varepsilon(s)\leqslant c^{\mathcal{S}_n,0}_{1-\gamma_n-\psi_n}\Big)=_{(4)}1-\gamma_n-\psi_n+o(1)=_{(5)}1-\gamma_n+o(1),
\end{aligned}
$$

where *(1)* is by the definition of $\mathcal{S}^R_n$, *(2)* is by the definition of $\varepsilon(s)$'s, *(3)* is by the fact that $c^{PI,0}_{1-\gamma_n-\psi_n}=c^{\mathcal{S}_n,0}_{1-\gamma_n-\psi_n}$, *(4)* is by Lemma A.5, and *(5)* is by the definition of ψ_n; see the statement of Lemma A.6. ■

A.2. Proofs of Theorems

Proof of Theorem 3.1. Let $\{\psi_n\}_{n\geqslant 1}$ be the sequence of positive numbers converging to zero defined in the statement of Lemma A.6, and let $\mathcal{S}^R_n$ be the subset of $\mathcal{S}_n$ defined in the

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

statement of Lemma A.7. Also, recall the random variable $\mathcal{V}_n$ defined in the beginning of this appendix and define

$$\underline{\mathcal{V}}_n = \min_{s \in \mathcal{S}_n} (V(s)/\widehat{V}(s))^{1/2}.$$

Then as in the proof of Lemma A.7, there exists a sequence of $\{w_n\}_{n \geqslant 1}$ of positive numbers converging to zero such that uniformly over $M \in \mathcal{M}_{\mathcal{H}}$,

$$\log p / n^{\kappa} \leqslant \omega_n \psi_n, \tag{A.11}$$

and, in addition,

$$\mathrm{P}_M \left(\left| c^{\mathcal{S}_n^R, 0}_{1-\alpha-\psi_n} (1/\mathcal{V}_n - 1) \right| > (\log p)^{1/2} / n^{\kappa} \right) = o(1), \tag{A.12}$$

and

$$\mathrm{P}_M \left(\left| c^{\mathcal{S}_n, 0}_{1-\alpha+\psi_n} (1/\underline{\mathcal{V}}_n - 1) \right| > (\log p)^{1/2} / n^{\kappa} \right) = o(1). \tag{A.13}$$

Furthermore, by Lemma A.5, uniformly over $M \in \mathcal{M}_{\mathcal{H}}$ and $\eta \in (0, 1)$,

$$\mathrm{P}_M \left(\max_{s \in \mathcal{S}_n^R} \varepsilon(s) \leqslant c^{\mathcal{S}_n^R, 0}_{\eta} \right) = \eta + o(1) \text{ and } \mathrm{P}_M \left(\max_{s \in \mathcal{S}_n} \varepsilon(s) \leqslant c^{\mathcal{S}_n, 0}_{\eta} \right) = \eta + o(1). \tag{A.14}$$

Finally, by Lemma A.6, uniformly over $M \in \mathcal{M}_{\mathcal{H}}$,

$$\mathrm{P}_M (c^{\mathcal{S}_n^R}_{1-\alpha} < c^{\mathcal{S}_n^R, 0}_{1-\alpha-\psi_m}) = o(1) \text{ and } \mathrm{P}_M (c^{\mathcal{S}_n}_{1-\alpha} > c^{\mathcal{S}_n, 0}_{1-\alpha+\psi_n}) = o(1). \tag{A.15}$$

Hence, uniformly over $M \in \mathcal{M}_{0,\mathcal{H}}$,

$$\begin{aligned}
\mathrm{P}_M (T \leqslant c^P_{1-\alpha}) &= \mathrm{P}_M \left(\max_{s \in \mathcal{S}_n} \textstyle\sum_{1 \leqslant i \leqslant n} w_i(s) Y_i / (\widehat{V}(s))^{1/2} \leqslant c^P_{1-\alpha} \right) \\
&\geqslant_{(1)} \mathrm{P}_M \left(\max_{s \in \mathcal{S}_n^R} \textstyle\sum_{1 \leqslant i \leqslant n} w_i(s) Y_i / (\widehat{V}(s))^{1/2} \leqslant c^P_{1-\alpha} \right) - \gamma_n + o(1) \\
&\geqslant_{(2)} \mathrm{P}_M \left(\max_{s \in \mathcal{S}_n^R} \textstyle\sum_{1 \leqslant i \leqslant n} w_i(s) Y_i / (\widehat{V}(s))^{1/2} \leqslant c^{\mathcal{S}_n^R}_{1-\alpha} \right) - 2\gamma_n + o(1) \\
&\geqslant_{(3)} \mathrm{P}_M \left(\max_{s \in \mathcal{S}_n^R} \textstyle\sum_{1 \leqslant i \leqslant n} w_i(s) \varepsilon_i / (\widehat{V}(s))^{1/2} \leqslant c^{\mathcal{S}_n^R}_{1-\alpha} \right) - 2\gamma_n + o(1) \\
&\geqslant_{(4)} \mathrm{P}_M \left(\mathcal{V}_n \max_{s \in \mathcal{S}_n^R} \varepsilon(s) \leqslant c^{\mathcal{S}_n^R}_{1-\alpha} \right) - 2\gamma_n + o(1) \\
&\geqslant_{(5)} \mathrm{P}_M \left(\mathcal{V}_n \max_{s \in \mathcal{S}_n^R} \varepsilon(s) \leqslant c^{\mathcal{S}_n^R, 0}_{1-\alpha-\psi_n} \right) - 2\gamma_n + o(1),
\end{aligned}$$

where *(1)* is by Lemma A.8 and the fact that $c^P_{1-\alpha} \geqslant 0$ for $\alpha < 1/2$, *(2)* is by Lemma A.7, *(3)* is by the fact that $\sum_{1 \leqslant i \leqslant n} w_i(s) f(X_i) \leqslant 0$ for all $s \in \mathcal{S}_n^R$ under $\mathcal{H}_0$, *(4)* is by the definition of $\mathcal{V}_n$ and the fact that $c^{\mathcal{S}_n^R}_{1-\alpha} \geqslant 0$ for $\alpha < 1/2$, and *(5)* is by (A.15). Furthermore,

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

uniformly over $M \in \mathcal{M}_{0,\mathcal{H}}$, for some constant $C' > 0$, the right-hand side of *(5)* is equal to

$$\begin{aligned}
&\mathrm{P}_M\Big(\max_{s\in\mathcal{S}_n^R} \varepsilon(s) \leqslant c^{\mathcal{S}_n^R,0}_{1-\alpha-\psi_n} + c^{\mathcal{S}_n^R,0}_{1-\alpha-\psi_n}(1/\mathcal{V}_n - 1)\Big) - 2\gamma_n + o(1)\\
&\quad \geqslant_{(6)} \mathrm{P}_M\Big(\max_{s\in\mathcal{S}_n^R} \varepsilon(s) \leqslant c^{\mathcal{S}_n^R,0}_{1-\alpha-\psi_n} - (\log p)^{1/2}/n^{\kappa}\Big) - 2\gamma_n + o(1)\\
&\quad \geqslant_{(7)} \mathrm{P}_M\Big(\max_{s\in\mathcal{S}_n^R} \varepsilon(s) \leqslant c^{\mathcal{S}_n^R,0}_{1-\alpha-\psi_n - C'\log p/n^{\kappa}}\Big) - 2\gamma_n + o(1)\\
&\quad \geqslant_{(8)} \mathrm{P}_M\Big(\max_{s\in\mathcal{S}_n^R} \varepsilon(s) \leqslant c^{\mathcal{S}_n^R,0}_{1-\alpha-\psi_n - C'\omega_n\psi_n}\Big) - 2\gamma_n + o(1)\\
&\quad =_{(9)} 1-\alpha-\psi_n - C'\omega_n\psi_n - 2\gamma_n + o(1) =_{(10)} 1-\alpha+o(1),
\end{aligned}$$

where *(6)* is by (A.12), *(7)* is by Lemma A.2, *(8)* is by (A.11), *(9)* is by (A.14), and *(10)* is by the definitions of ψ_n, ω_n, and γ_n. The first asserted claim follows.

To prove the second asserted claim, note that uniformly over $M \in \mathcal{M}_{00,\mathcal{H}}$,

$$\begin{aligned}
\mathrm{P}(T \leqslant c^P_{1-\alpha}) &=_{(11)} \mathrm{P}\Big(\max_{s\in\mathcal{S}_n} \textstyle\sum_{1\leqslant i\leqslant n} w_i(s)\varepsilon_i/(\widehat{V}(s))^{1/2} \leqslant c^P_{1-\alpha}\Big)\\
&\leqslant_{(12)} \mathrm{P}\Big(\max_{s\in\mathcal{S}_n} \textstyle\sum_{1\leqslant i\leqslant n} w_i(s)\varepsilon_i/(\widehat{V}(s))^{1/2} \leqslant c^{\mathcal{S}_n}_{1-\alpha}\Big)\\
&\leqslant_{(13)} \mathrm{P}\Big(\underline{\mathcal{V}}_n \max_{s\in\mathcal{S}_n} \varepsilon(s) \leqslant c^{\mathcal{S}_n}_{1-\alpha}\Big) + o(1)\\
&\leqslant_{(14)} \mathrm{P}\Big(\underline{\mathcal{V}}_n \max_{s\in\mathcal{S}_n} \varepsilon(s) \leqslant c^{\mathcal{S}_n,0}_{1-\alpha+\psi_n}\Big) + o(1),
\end{aligned}$$

where *(11)* is by the fact that $\sum_{1\leqslant i\leqslant n} w_i(s) f(X_i) = 0$ for all $s \in \mathcal{S}_n$ whenever $f(\cdot) \equiv C$ for some constant C, *(12)* is by the fact that $c^{SD}_{1-\alpha} \leqslant c^{OS}_{1-\alpha} \leqslant c^{PI}_{1-\alpha} = c^{\mathcal{S}_n}_{1-\alpha}$, *(13)* is by the definition of $\underline{\mathcal{V}}_n$ and the fact that $c^{\mathcal{S}_n}_{1-\alpha} \geqslant 0$ for $\alpha < 1/2$, and *(14)* is by (A.15). Furthermore, uniformly over $M \in \mathcal{M}_{00,\mathcal{H}}$, the right-hand side of *(14)* is equal to

$$\begin{aligned}
&\mathrm{P}\Big(\max_{s\in\mathcal{S}_n} \varepsilon(s) \leqslant c^{\mathcal{S}_n,0}_{1-\alpha+\psi_n} + c^{\mathcal{S}_n,0}_{1-\alpha+\psi_n}(1/\underline{\mathcal{V}}_n - 1)\Big) + o(1)\\
&\quad \leqslant_{(15)} \mathrm{P}\Big(\max_{s\in\mathcal{S}_n} \varepsilon(s) \leqslant c^{\mathcal{S}_n,0}_{1-\alpha+\psi_n} + (\log p)^{1/2}/n^{\kappa}\Big) + o(1)\\
&\quad \leqslant_{(16)} \mathrm{P}\Big(\max_{s\in\mathcal{S}_n} \varepsilon(s) \leqslant c^{\mathcal{S}_n,0}_{1-\alpha+\psi_n + C'\log p/n^{\kappa}}\Big) + o(1)\\
&\quad \leqslant_{(17)} \mathrm{P}\Big(\max_{s\in\mathcal{S}_n} \varepsilon(s) \leqslant c^{\mathcal{S}_n,0}_{1-\alpha+\psi_n + C'\omega_n\psi_n}\Big) + o(1)\\
&\quad =_{(18)} 1-\alpha+\psi_n + C'\omega_n\psi_n + o(1) \leqslant_{(19)} 1-\alpha+o(1),
\end{aligned}$$

where *(15)* is by (A.13), *(16)* is by Lemma A.2, *(17)* is by (A.11), *(18)* is by (A.14), and *(19)* is by the definition of ψ_n and ω_n. The second asserted claim follows. ■

Proof of Theorem 3.2. Let $\{\psi_n\}_{n\geqslant 1}$ be the sequence of positive numbers converging to zero defined in the statement of Lemma A.6. Also, recall the random variables $w_i(s)$ and $\varepsilon(s)$ defined in the beginning of this appendix.

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

Let $x_1, x_2 \in \mathcal{X}$ be such that $x_1 < x_2$ but $f(x_1) > f(x_2)$. By the mean value theorem, there exists $x_0 \in (x_1, x_2)$ such that

$$f'(x_0)(x_2 - x_1) = f(x_2) - f(x_1) < 0.$$

Therefore, $f'(x_0) < 0$, and since $f'(\cdot)$ is continuous, it follows that there exists $\Delta_x > 0$ such that $f'(x) < f'(x_0)/2$ for all $x \in [x_0 - \Delta_x, x_0 + \Delta_x]$. Next, applying Assumption H4 to the interval $[x_0 - \Delta_x, x_0 + \Delta_x]$ shows that there exists an event $\mathcal{A}_n$ such that $\mathrm{P}_M(\mathcal{A}_n) = 1 - o(1)$ and whenever $\mathcal{A}_n$ holds, there exists $s \in \mathcal{S}_n$ that satisfies conditions (i)–(iii) of Assumption H4. Let $\bar{s}_n$ be an element of $\mathcal{S}_n$ that satisfies these conditions when $\mathcal{A}_n$ holds and that is defined arbitrarily when $\mathcal{A}_n$ does not hold. Then on $\mathcal{A}_n$, by Assumption L1 and condition (ii) of Assumption H4,

$$V(\bar{s}_n) = \sum_{1 \leqslant i \leqslant n} \sigma_i^2 \left(\sum_{1 \leqslant j \leqslant n} \mathrm{sign}(X_j - X_i) Q(X_i, X_j, \bar{s}_n) \right)^2 \leqslant Cn^3 \tag{A.16}$$

for some constant $C > 0$. In addition, by Assumption L2 and Lemma C.1 in Appendix A.2, there exists an event $\mathcal{B}_n$ such that $\mathrm{P}_M(\mathcal{B}_n) = 1 - o(1)$ and on $\mathcal{B}_n$, for all $x_1, x_2 \in \mathcal{X}$ with $x_2 - x_1 \geqslant h_n$,

$$(c_3/2)n(x_2 - x_1) \leqslant \left| \{i = 1, \ldots, n \colon X_i \in [x_1, x_2]\} \right| \leqslant (3C_3/2)n(x_2 - x_1).$$

Hence, for some constant $c > 0$, on $\mathcal{A}_n \cap \mathcal{B}_n$, for the intervals $[x_{l1}, x_{r1}]$ and $[x_{l2}, x_{r2}]$ appearing in condition (iii) of Assumption H4 and corresponding to $\bar{s}_n$,

$$\left| \{i = 1, \ldots, n : X_i \in [x_{l1}, x_{r1}]\} \right| \geqslant cn \text{ and } \left| \{i = 1, \ldots, n : X_i \in [x_{l2}, x_{r2}]\} \right| \geqslant cn.$$

Then on $\mathcal{A}_n \cap \mathcal{B}_n$, by conditions (i) and (iii) of Assumption H4,

$$\sum_{1 \leqslant i \leqslant n} w_i(\bar{s}_n) f(X_i) = \frac{1}{2} \sum_{1 \leqslant i, j \leqslant n} (f(X_i) - f(X_j)) \mathrm{sign}(X_j - X_i) Q(X_i, X_j, \bar{s}_n) \geqslant cn^2 \tag{A.17}$$

for some constant $c > 0$. Combining (A.16) and (A.17) shows that on $\mathcal{A}_n \cap \mathcal{B}_n$,

$$\sum_{1 \leqslant i \leqslant n} \frac{w_i(\bar{s}_n) f(X_i)}{(V(\bar{s}_n))^{1/2}} \geqslant cn^{1/2} \tag{A.18}$$

for some constant $c > 0$. On the other hand, $\log p = o(n)$ by Assumption H3. Hence, given that $\mathrm{P}_M(\mathcal{A}_n \cap \mathcal{B}_n) = 1 - o(1)$, it follows that there exists a sequence $\{\omega_n\}_{n \geqslant 1}$ of positive numbers such that $\omega_n \to \infty$ as $n \to \infty$ and

$$\mathrm{P}_M \left(\omega_n (\log p)^{1/2} > \sum_{1 \leqslant i \leqslant n} \frac{w_i(\bar{s}_n) f(X_i)}{(V(\bar{s}_n))^{1/2}} \right) = o(1). \tag{A.19}$$

Furthermore, let $\eta_n = 1 - 2C/\omega_n$ where C is the same constant as that in (A.2) in the statement of Lemma A.3. Then by Lemma A.3, $c_{\eta_n}^{\mathcal{S}_n, 0} \leqslant (\omega_n/2)(\log p)^{1/2}$ and by Lemma A.5,

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

$$\mathrm{P}_M\left(\max_{s\in\mathcal{S}_n}(-\varepsilon(s)) > (\omega_n/2)(\log p)^{1/2}\right) \leqslant \mathrm{P}_M\left(\max_{s\in\mathcal{S}_n}(-\varepsilon(s)) > c_{\eta_n}^{\mathcal{S}_n,0}\right) = 1-\eta_n+o(1) = o(1). \tag{A.20}$$

Finally, by Lemma A.6,

$$\mathrm{P}_M(c_{1-\alpha}^{\mathcal{S}_n} > c_{1-\alpha+\psi_n}^{\mathcal{S}_n,0}) = o(1). \tag{A.21}$$

Hence,

$$\begin{aligned}
\mathrm{P}_M(T \leqslant c_{1-\alpha}^{P}) &\leqslant_{(1)} \mathrm{P}_M(T \leqslant c_{1-\alpha}^{\mathcal{S}_n}) \\
&\leqslant_{(2)} \mathrm{P}_M\left(\textstyle\sum_{1\leqslant i\leqslant n} w_i(\bar{s}_n)Y_i/(\widehat{V}(\bar{s}_n))^{1/2} \leqslant c_{1-\alpha}^{\mathcal{S}_n}\right) \\
&=_{(3)} \mathrm{P}_M\left(\textstyle\sum_{1\leqslant i\leqslant n} w_i(\bar{s}_n)Y_i/(V(\bar{s}_n))^{1/2} \leqslant c_{1-\alpha}^{\mathcal{S}_n}(\widehat{V}(\bar{s}_n)/V(\bar{s}_n))^{1/2}\right) \\
&\leqslant_{(4)} \mathrm{P}_M\left(\textstyle\sum_{1\leqslant i\leqslant n} w_i(\bar{s}_n)Y_i/(V(\bar{s}_n))^{1/2} \leqslant 2c_{1-\alpha}^{\mathcal{S}_n}\right)+o(1) \\
&\leqslant_{(5)} \mathrm{P}_M\left(\textstyle\sum_{1\leqslant i\leqslant n} w_i(\bar{s}_n)Y_i/(V(\bar{s}_n))^{1/2} \leqslant 2c_{1-\alpha+\psi_n}^{\mathcal{S}_n,0}\right)+o(1),
\end{aligned}$$

where *(1)* is by the fact that $c_{1-\alpha}^{SD} \leqslant c_{1-\alpha}^{OS} \leqslant c_{1-\alpha}^{PI} = c_{1-\alpha}^{\mathcal{S}_n}$, *(2)* is by the definition of the test statistic T, *(3)* is by a rearrangement, *(4)* is by Assumption H2, and *(5)* is by (A.21). Furthermore, by Lemma A.3, for some constant $C > 0$, possibly depending on α, the right-hand side of *(5)* is bounded from above by

$$\begin{aligned}
&\mathrm{P}_M\left(\textstyle\sum_{1\leqslant i\leqslant n} w_i(\bar{s}_n)Y_i/(V(\bar{s}_n))^{1/2} \leqslant C(\log p)^{1/2}\right)+o(1) \\
&\quad\leqslant_{(6)} \mathrm{P}_M\left(\varepsilon(\bar{s}_n)+\textstyle\sum_{1\leqslant i\leqslant n} w_i(\bar{s}_n)f(X_i)/(V(\bar{s}_n))^{1/2} \leqslant C(\log p)^{1/2}\right)+o(1) \\
&\quad\leqslant_{(7)} \mathrm{P}_M\left(\textstyle\sum_{1\leqslant i\leqslant n} w_i(\bar{s}_n)f(X_i)/(V(\bar{s}_n))^{1/2} \leqslant (C+\omega_n/2)(\log p)^{1/2}\right)+o(1) \\
&\quad\leqslant_{(8)} \mathrm{P}_M\left(\omega_n(\log p)^{1/2} \leqslant (C+\omega_n/2)(\log p)^{1/2}\right)+o(1) =_{(9)} o(1),
\end{aligned}$$

where *(6)* is by the definition of $\varepsilon(\bar{s}_n)$, *(7)* by (A.20), *(8)* is by (A.19), and *(9)* is by the fact that $\omega_n \to \infty$ as $n \to \infty$. The asserted claim follows. ■

Proof of Theorem 3.3. The proof is closely related to that of Theorem 3.2 with $\mathrm{P}_M(\cdot)$ replaced by $\mathrm{P}_{M_n}(\cdot)$. Note that since the sequence $\{M_n\}_{n\geqslant 1}$ is such that $M_n \in \mathcal{M}_{\mathcal{H}}$ for all $n \geqslant 1$, the results of Lemmas A.3, A.5, and A.6, which were used in the proof of Theorem 3.2, hold under the sequence of models $\{M_n\}_{n\geqslant 1}$. Then, it follows from the same arguments as those used in the proof of Theorem 3.2 that the bounds (A.17) and (A.18) become

$$\sum_{1\leqslant i\leqslant n} w_i(\bar{s}_n)f(X_i) = \frac{1}{2}\sum_{1\leqslant i,j\leqslant n}(f(X_i)-f(X_j))\mathrm{sign}(X_j-X_i)Q(X_i,X_j,\bar{s}_n) \geqslant C\ell_n n^2$$

and

$$\sum_{1\leqslant i\leqslant n}\frac{w_i(\bar{s}_n)f(X_i)}{(V(\bar{s}_n))^{1/2}} \geqslant c\ell_n n^{1/2},$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

respectively, and so (A.19) continues to hold for some sequence $\{\omega_n\}_{n\geqslant 1}$ of positive numbers such that $\omega_n \to \infty$ as $n \to \infty$ as long as $(\log p/n)^{1/2} = o(l_n)$. The rest of the proof is the same as that in the proof of Theorem 3.2. ■

Proof of Theorem 3.4. The asserted claim follows if

$$\mathrm{P}_{M_n}(T \leqslant c_{1-\alpha}^P) \to 0 \text{ for any } \{M_n\}_{n\geqslant 1} \subset \mathcal{M}_{\mathcal{H}} \text{ such that } M_n \in \mathcal{M}_{\mathcal{H},l_n} \text{ for all } n \geqslant 1. \tag{A.22}$$

To prove (A.22), take any sequence $\{M_n\}_{n\geqslant 1} \subset \mathcal{M}_{\mathcal{H}}$ of models such that $M_n \in \mathcal{M}_{\mathcal{H},l_n}$ for all $n \geqslant 1$ and apply arguments similar to those used in the proof of Theorem 3.2. Specifically, since $\inf_{x\in\mathcal{X}} f'_{M_n}(x) < -l_n$ and $h_n^\beta = (\log p/n)^{\beta/(2\beta+3)} = o(l_n)$, it follows from Assumption L3 that for all sufficiently large n, there exists an interval $[x_{n,1}, x_{n,2}] \subset \mathcal{X}$ such that $|x_{n,2} - x_{n,1}| = h_n$ and $f'(x) < -l_n/2$ for all $x \in [x_{n,1}, x_{n,2}]$. Next, using the same arguments as those in the proof of Theorem 3.2 but applying Assumption H4 to the interval $[x_{n,1}, x_{n,2}]$ instead of the interval $[x_0 - \Delta_x, x_0 + \Delta_x]$, it follows that there exist events $\mathcal{A}_n$ and $\mathcal{B}_n$ and a weighting function $\bar{s}_n \in \mathcal{S}_n$ such that $\mathrm{P}_{M_n}(\mathcal{A}_n \cap \mathcal{B}_n) = 1 - o(1)$ and on $\mathcal{A}_n \cap \mathcal{B}_n$,

$$cnh_n \leqslant \Big|\{i = 1, \ldots, n\colon X_i \in [x_{n,1}, x_{n,2}]\}\Big| \leqslant Cnh_n,$$

$$V(\bar{s}_n) = \sum_{1\leqslant i\leqslant n} \sigma_i^2 \left(\sum_{1\leqslant j\leqslant n} \mathrm{sign}(X_j - X_i) Q(X_i, X_j, \bar{s}_n)\right)^2 \leqslant C'(nh_n)^3 h_n^{2k},$$

$$\sum_{1\leqslant i\leqslant n} w_i(\bar{s}_n) f(X_i) = \frac{1}{2} \sum_{1\leqslant i,j\leqslant n} (f(X_i) - f(X_j))\mathrm{sign}(X_j - X_i) Q(X_i, X_j, \bar{s}_n) \geqslant c'(l_n h_n)(nh_n)^2 h_n^k,$$

and

$$\sum_{1\leqslant i\leqslant n} \frac{w_i(\bar{s}_n) f(X_i)}{(V(\bar{s}_n))^{1/2}} \geqslant c''(l_n h_n)(nh_n)^{1/2} = c'' l_n n^{1/2} h_n^{3/2} = c'' l_n n^{1/2} \left(\frac{\log p}{n}\right)^{3/(4\beta+6)}$$

for some strictly positive constants c, C, c', C', and c''. Hence, (A.19) continues to hold for some sequence $\{\omega_n\}_{n\geqslant 1}$ of positive numbers such that $\omega_n \to \infty$ as $n \to \infty$ as long as $(\log p/n)^{\beta/(2\beta+3)} = o(l_n)$. The rest of the proof is the same as that in the proof of Theorem 3.2. ■

Proof of Theorem 3.5. Without loss of generality, I can and will assume that $c_2 = 0$ and $C_2 = 1$, so that $\mathcal{X} = [0, 1]$. Also, let $\bar{h}_n = (c^2 \log n/n)^{1/(2\beta+3)}$ for sufficiently small constant $c > 0$. In addition, let $L_n = [1/(4\bar{h}_n)]$ where $[x]$ is the largest integer smaller than or equal to x. Moreover, for $l = 1, \ldots, L_n$, let $x_{n,l} = 4\bar{h}_n(l-1)$ and define $f_{n,l}\colon \mathcal{X} \to \mathbb{R}$ by $f_{n,l}(0) = 0$ and

$$f'_{n,l}(x) = \begin{cases} 0 & \text{if } x \leqslant x_{n,l}, \\ -L(x - x_{n,l})^\beta & \text{if } x \in (x_{n,l}, x_{n,l} + \bar{h}_n], \\ -L(x_{n,l} + 2\bar{h}_n - x)^\beta & \text{if } x \in (x_{n,l} + \bar{h}_n, x_{n,l} + 2\bar{h}_n], \\ L(x - x_{n,l} - 2\bar{h}_n)^\beta & \text{if } x \in (x_{n,l} + 2\bar{h}_n, x_{n,l} + 3\bar{h}_n], \\ L(x_{n,l} + 4\bar{h}_n - x)^\beta & \text{if } x \in (x_{n,l} + 3\bar{h}_n, x_{n,l} + 4\bar{h}_n], \\ 0 & \text{if } x > x_{n,l} + 4\bar{h}_n. \end{cases}$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

Finally, let $f_{n,0}\colon \mathcal{X} \to \mathbb{R}$ be the function given by $f_{n,0}(\cdot) \equiv 0$.

Now, for $l = 0, \ldots, L_n$, consider the model $M_{n,l}$ with the regression function $f_l(\cdot)$, X distributed uniformly over $[0,1]$, and ε distributed as a $N(0,1)$ random variable independently of X. Note that $M_{n,0}$ belongs to $\mathcal{M}_{\mathcal{L}}$ and satisfies $\mathcal{H}_0$. In addition, for $l = 1, \ldots, L_n$, the model $M_{n,l}$ belongs to $\mathcal{M}_{\mathcal{L}}$, does not satisfy $\mathcal{H}_0$, and, moreover, has $\inf_{x \in \mathcal{X}} f'_{n,l}(x) = -L\bar{h}_n^{\beta} = -L(c^2 \log n/n)^{\beta/(2\beta+3)}$.

Next, consider any test $\psi = \psi(\{X_i, Y_i\}_{1 \leqslant i \leqslant n})$ such that $\mathrm{E}_{M_{n,0}}[\psi] \leqslant \alpha + o(1)$. Then following the argument from Dumbgen and Spokoiny (2001) gives

$$\begin{aligned}
\inf_{M \in \mathcal{M}_{\mathcal{L}}} \mathrm{E}_M[\psi] - \alpha &\leqslant \min_{1 \leqslant l \leqslant L_n} \mathrm{E}_{M_{n,l}}[\psi] - \mathrm{E}_{M_{n,0}}[\psi] + o(1) \\
&\leqslant \textstyle\sum_{1 \leqslant l \leqslant L_n} \mathrm{E}_{M_{n,l}}[\psi]/L_n - \mathrm{E}_{M_{n,0}}[\psi] + o(1) \\
&= \textstyle\sum_{1 \leqslant l \leqslant L_n} \mathrm{E}_{M_{n,0}}[\psi \rho_{n,l}]/L_n - \mathrm{E}_{M_{n,0}}[\psi] + o(1) \\
&= \textstyle\sum_{1 \leqslant l \leqslant L_n} \mathrm{E}_{M_{n,0}}[\psi(\rho_{n,l} - 1)]/L_n + o(1) \\
&\leqslant \mathrm{E}_{M_{n,0}}\Big[\psi \Big| \textstyle\sum_{1 \leqslant l \leqslant L_n} \rho_{n,l}/L_n - 1\Big|\Big] + o(1) \\
&\leqslant \mathrm{E}_{M_{n,0}}\Big[\Big| \textstyle\sum_{1 \leqslant l \leqslant L_n} \rho_{n,l}/L_n - 1\Big|\Big] + o(1),
\end{aligned}$$

where $\rho_{n,l}$ is the likelihood ratio of observing $\{X_i, Y_i\}_{1 \leqslant i \leqslant n}$ under the models $M_{n,l}$ and $M_{n,0}$. Furthermore,

$$\rho_{n,l} = \exp\left(\sum_{1 \leqslant i \leqslant n} Y_i f_l(X_i) - \sum_{1 \leqslant i \leqslant n} f_l(X_i)^2/2\right) = \exp(\omega_{n,l}\xi_{n,l} - \omega_{n,l}^2/2),$$

where

$$\omega_{n,l} = \left(\sum_{1 \leqslant i \leqslant n} f_l(X_i)^2\right)^{1/2} \text{ and } \xi_{n,l} = \sum_{1 \leqslant i \leqslant n} Y_i f_l(X_i)/\omega_{n,l}.$$

Note that in the model $M_{n,0}$, conditional on $\{X_i\}_{1 \leqslant i \leqslant n}$, each $\xi_{n,l}$ is a $N(0,1)$ random variable, so that

$$\mathrm{E}_{M_{n,0}}\Big[|\rho_{n,l}||\{X_i\}_{1 \leqslant i \leqslant n}\Big] = \mathrm{E}_{M_{n,0}}\Big[\rho_{n,l}|\{X_i\}_{1 \leqslant i \leqslant n}\Big] = 1$$

for all $l = 1, \ldots, L_n$, and so

$$\mathrm{E}_{M_{n,0}}\left[\left|\frac{1}{L_n}\sum_{1 \leqslant l \leqslant L_n} \rho_{n,l} - 1\right| |\{X_i\}_{1 \leqslant i \leqslant n}\right] = \frac{1}{L_n}\sum_{1 \leqslant l \leqslant L_n} \mathrm{E}_{M_{n,0}}\Big[|\rho_{n,l}||\{X_i\}_{1 \leqslant i \leqslant n}\Big] + 1 \leqslant 2. \tag{A.23}$$

In addition, by construction of the functions $f_{n,l}(\cdot)$ and Lemma C.1, in the model $M_{n,0}$,

$$\omega_{n,l} \leqslant C(n\bar{h}_n)^{1/2}\bar{h}_n^{1+\beta} = Cn^{1/2}\bar{h}_n^{3/2+\beta} = Cn^{1/2}\left(\frac{c^2 \log n}{n}\right)^{1/2} = cC(\log n)^{1/2}$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

for all $l = 1, \ldots, L_n$ with probability $1 - o(1)$ for some constant $C > 0$ where c is the same constant as that in the definition of $\bar{h}_n$. Let $\mathcal{A}_n$ denote the event that $\omega_{n,l} \leqslant cC(\log n)^{1/2}$ for all $l = 1, \ldots, L_n$. Then $\mathrm{P}_{M_{n,0}}(\mathcal{A}_n) \to 1$ and by (A.23),

$$\mathrm{E}_{M_{n,0}}\left[\left|\frac{1}{L_n}\sum_{1\leqslant l\leqslant L_n}\rho_{n,l} - 1\right|\right] \leqslant \mathrm{E}_{M_{n,0}}\left[\left|\frac{1}{L_n}\sum_{1\leqslant l\leqslant L_n}\rho_{n,l} - 1\right| |\mathcal{A}_n\right] + 2(1 - \mathrm{P}_{M_{n,0}}(\mathcal{A}_n)). \tag{A.24}$$

Furthermore, on $\mathcal{A}_n$, since in the model $M_{n,0}$, conditional on $\{X_i\}_{1\leqslant i\leqslant n}$, $\{\xi_{n,l}\}_{1\leqslant l\leqslant L_n}$ are independent $N(0,1)$ random variables,

$$\begin{aligned}
&\mathrm{E}_{M_{n,0}}\left[\left|\textstyle\sum_{1\leqslant l\leqslant L_n}\rho_{n,l}/L_n - 1\right| |\{X_i\}_{1\leqslant i\leqslant n}\right]^2 \\
&\quad\leqslant \mathrm{E}_{M_{n,0}}\left[\left(\textstyle\sum_{1\leqslant l\leqslant L_n}\rho_{n,l}/L_n - 1\right)^2 |\{X_i\}_{1\leqslant i\leqslant n}\right] \\
&\quad\leqslant \textstyle\sum_{1\leqslant l\leqslant L_n}\mathrm{E}_{M_{n,0}}\left[\rho_{n,l}^2/L_n^2 |\{X_i\}_{1\leqslant i\leqslant n}\right] \\
&\quad\leqslant \textstyle\sum_{1\leqslant l\leqslant L_n}\mathrm{E}_{M_{n,0}}\left[\exp(2\omega_{n,l}\xi_{n,l} - \omega_{n,l}^2)/L_n^2 |\{X_i\}_{1\leqslant i\leqslant n}\right] \\
&\quad\leqslant \textstyle\sum_{1\leqslant l\leqslant L_n}\exp(\omega_{n,l}^2)/L_n^2 \leqslant \max_{1\leqslant l\leqslant L_n}\exp(\omega_{n,l}^2)/L_n \\
&\quad\leqslant \exp\left(c^2C^2\log n - \log L_n\right) = o(1)
\end{aligned}$$

because the constant c in the last line is arbitrarily small and $\log n \leqslant C\log L_n$ for some constant $C > 0$. Combining the last bound with (A.24) gives $\inf_{M\in\mathcal{M}_{\mathcal{L}}}\mathrm{E}_M[\psi] \leqslant \alpha + o(1)$, and completes the proof of the theorem. ■

APPENDIX B: Proofs for Section 4

Proof of Theorem 4.1. Throughout the proof, I will assume that the observations (X_i, Y_i) are ordered so that $X_i \leqslant X_j$ whenever $i < j$. Since all the arguments are conditional on $\{X_i\}_{1\leqslant i\leqslant n}$, this assumption is without loss of generality. In addition, I will use C to denote a strictly positive constant that can change from place to place.

Note that since $\sigma_i \geqslant c_1$ for all $i = 1, \ldots, n$ by Assumption L1, it follows that

$$|\widehat{\sigma}_i - \sigma_i| = \left|\frac{\widehat{\sigma}_i^2 - \sigma_i^2}{\widehat{\sigma}_i + \sigma_i}\right| \leqslant \frac{|\widehat{\sigma}_i^2 - \sigma_i^2|}{c_1}, \quad \text{for all } i = 1, \ldots, n,$$

and so it suffices to bound $\max_{1\leqslant i\leqslant n}|\widehat{\sigma}_i^2 - \sigma_i^2|$. In addition, note that by Assumptions L1 and L3,

$$|\sigma_j^2 - \sigma_i^2| = |\sigma_j + \sigma_i|\cdot|\sigma_j - \sigma_i| \leqslant 2C_1L|X_j - X_i|, \quad \text{for all } i, j = 1, \ldots, n.$$

Moreover, note that

$$\widehat{\sigma}_i^2 - \sigma_i^2 = \frac{1}{2|J(i)|}\sum_{j\in J(i)'}(Y_{j+1} - Y_j)^2 - \sigma_i^2, \quad \text{for all } i = 1, \ldots, n$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

where $J(i)' = \{j \in J(i) : j+1 \in J(i)\}$. Therefore,

$$\left|(\widehat{\sigma}_i^2 - \sigma_i^2) - (p_{i,1} + p_{i,2} + p_{i,3} - p_{i,4})\right| \leqslant 2C_1 L b_n + \frac{(Lb_n)^2}{2} + \frac{C_1^2}{|J(i)|}, \quad \text{for all } i = 1, \ldots, n$$

by Assumptions L1 and L3 where

$$p_{i,1} = \frac{1}{2|J(i)|} \sum_{j \in J(i)'} (\varepsilon_{j+1}^2 - \sigma_{j+1}^2), \qquad p_{i,2} = \frac{1}{2|J(i)|} \sum_{j \in J(i)'} (\varepsilon_j^2 - \sigma_j^2),$$

$$p_{i,3} = \frac{1}{|J(i)|} \sum_{j \in J(i)'} (f(X_{j+1}) - f(X_j))(\varepsilon_{j+1} - \varepsilon_j), \quad p_{i,4} = \frac{1}{|J(i)|} \sum_{j \in J(i)'} \varepsilon_j \varepsilon_{j+1}.$$

In the rest of the proof, I derive bounds on $J(i)$, $p_{i,1}$, $p_{i,2}$, $p_{i,3}$, and $p_{i,4}$ that hold uniformly over $i = 1, \ldots, n$.

Since $b_n = C_b(\log n)^{1/2}/n^{1/4}$, by Assumption L2 and Lemma C.1 in Appendix A.2, the event $\mathcal{A}_n$ that

$$c_3 b_n n \leqslant |J(i)| \leqslant 3C_3 b_n n, \quad \text{for all } i = 1, \ldots, n$$

satisfies $\mathrm{P}_M(\mathcal{A}_n) = 1 - o(1)$ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$. Note also that $J(i)$'s depend only on $\{X_i\}_{1 \leqslant i \leqslant n}$. Therefore, applying Lemma C.2 conditional on $\{X_i\}_{1 \leqslant i \leqslant n}$ shows that on $\mathcal{A}_n$,

$$\begin{aligned} \mathrm{E}_M\left[\max_{1 \leqslant i \leqslant n} |p_{i,1}| \,|\, \{X_i\}_{1 \leqslant i \leqslant n}\right] &\leqslant C\left(\frac{\sqrt{b_n n \log n}}{b_n n} + \frac{n^{1/2} \log n}{b_n n}\right) \\ &= C\left(\sqrt{\frac{\log n}{b_n n}} + \frac{\log n}{b_n n^{1/2}}\right) \leqslant \frac{C \log n}{b_n n^{1/2}} \end{aligned}$$

uniformly over $M \in \mathcal{M}_{\mathcal{L}}$ since

$$\max_{1 \leqslant l \leqslant n} \sum_{j \in J(l)'} \mathrm{E}_M\left[\varepsilon_{j+1}^4 \,|\, \{X_i\}_{1 \leqslant i \leqslant n}\right] \leqslant 3C_1^4 C_3 b_n n$$

and

$$\left(\mathrm{E}_M\left[\max_{1 \leqslant l \leqslant n} \varepsilon_l^4 \,|\, \{X_i\}_{1 \leqslant i \leqslant n}\right]\right)^{1/2} \leqslant \left(\mathrm{E}_M\left[\sum_{1 \leqslant l \leqslant n} \varepsilon_l^4 \,|\, \{X_i\}_{1 \leqslant i \leqslant n}\right]\right)^{1/2} \leqslant C_1^2 n^{1/2}$$

for all $M \in \mathcal{M}_{\mathcal{L}}$ by Assumption L1. Similarly, on $\mathcal{A}_n$,

$$\mathrm{E}_M\left[\max_{1 \leqslant i \leqslant n} |p_{i,2}| \,|\, \{X_i\}_{1 \leqslant i \leqslant n}\right] \leqslant \frac{C \log n}{b_n n^{1/2}}$$

uniformly over $M \in \mathcal{M}_{\mathcal{L}}$. Furthermore, applying Lemma C.2 conditional on $\{X_i\}_{1 \leqslant i \leqslant n}$ again and using similar calculations shows that on $\mathcal{A}_n$,

$$\mathrm{E}_M\left[\frac{1}{|J(i)|}\left|\sum_{j \in J(i)' : j \text{ odd}} \varepsilon_j \varepsilon_{j+1}\right| \,\middle|\, \{X_i\}_{1 \leqslant i \leqslant n}\right] \leqslant \frac{C \log n}{b_n n^{1/2}}$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

and

$$\mathrm{E}_M\left[\frac{1}{|J(i)|}\left|\sum_{j\in J(i)':\, j \text{ even}} \varepsilon_j\varepsilon_{j+1}\right| \{X_i\}_{1\leqslant i\leqslant n}\right] \leqslant \frac{C\log n}{b_n n^{1/2}}$$

so that

$$\mathrm{E}_M\left[\max_{1\leqslant i\leqslant n} |p_{i,4}| |\{X_i\}_{1\leqslant i\leqslant n}\right] \leqslant \frac{C\log n}{b_n n^{1/2}}$$

uniformly over $M \in \mathcal{M}_{\mathcal{L}}$. Moreover, applying Lemma C.2 conditional on $\{X_i\}_{1\leqslant i\leqslant n}$ one more time shows that on $\mathcal{A}_n$,

$$\begin{aligned}&\mathrm{E}_M\left[\frac{1}{|J(i)|}\left|\sum_{j\in J(i)'} (f(X_{j+1})-f(X_j))\varepsilon_{j+1}\right| \{X_i\}_{1\leqslant i\leqslant n}\right]\\ &\quad\leqslant C\left(\frac{b_n(b_n n)^{1/2}(\log n)^{1/2}}{b_n n} + \frac{b_n n^{1/4}\log n}{b_n n}\right) \leqslant \frac{C\log n}{b_n n^{1/2}}\end{aligned}$$

uniformly over $M \in \mathcal{M}_{\mathcal{L}}$ since

$$\max_{1\leqslant l\leqslant n} \sum_{j\in J(l)'} \mathrm{E}_M\Big[(f(X_{j+1})-f(X_j))^2\varepsilon_{j+1}^2 |\{X_i\}_{1\leqslant i\leqslant n}\Big] \leqslant C_1^2\cdot(Lb_n)^2\cdot(3C_3 b_n n)$$

and

$$\begin{aligned}&\left(\mathrm{E}_M\Big[\max_{1\leqslant l\leqslant n}\max_{j\in J(l)'}(f(X_{j+1}-f(X_j))^2\varepsilon_{j+1}^2 |\{X_i\}_{1\leqslant i\leqslant n}\Big]\right)^{1/2}\\ &\quad\leqslant (Lb_n)\cdot\left(\mathrm{E}_M\Big[\max_{1\leqslant i\leqslant n}\varepsilon_i^2 |\{X_i\}_{1\leqslant i\leqslant n}\Big]\right)^{1/2} \leqslant (Lb_n)\cdot(C_1 n^{1/4})\end{aligned}$$

for all $M \in \mathcal{M}_{\mathcal{L}}$ by Assumption L1. Similarly,

$$\mathrm{E}_M\left[\frac{1}{|J(i)|}\left|\sum_{j\in J(i)'} (f(X_{j+1})-f(X_j))\varepsilon_j\right| |\{X_i\}_{1\leqslant i\leqslant n}\right] \leqslant \frac{C\log n}{b_n n^{1/2}},$$

and so

$$\mathrm{E}_M\left[\max_{1\leqslant i\leqslant n} |p_{i,3}| |\{X_i\}_{1\leqslant i\leqslant n}\right] \leqslant \frac{C\log n}{b_n n^{1/2}}$$

uniformly over $M \in \mathcal{M}_{\mathcal{L}}$. Combining presented bounds shows that on $\mathcal{A}_n$,

$$\mathrm{E}_M\left[\max_{1\leqslant i\leqslant n} |\widehat{\sigma}_i^2-\sigma_i^2| |\{X_i\}_{1\leqslant i\leqslant n}\right] \leqslant C\left(b_n + \frac{1}{b_n n} + \frac{\log n}{b_n n^{1/2}}\right) \leqslant \frac{C(\log n)^{1/2}}{n^{1/4}}$$

uniformly over $M \in \mathcal{M}_{\mathcal{L}}$. The asserted claim now follows from Markov's inequality since $\mathrm{P}_M(\mathcal{A}_n) = 1-o(1)$ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$. ■

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

Proof of Theorem 4.2. Fix $\kappa \in (0, 1/4)$ and note that by Theorem 4.1, Assumption H1 holds uniformly over $M \in \mathcal{M}_{\mathcal{L}}$ for this κ. In this proof, I will verify Assumptions H2 and H3 with this κ (and Assumption H4 does not depend on κ).

I first verify Assumption H3. Note that since $\mathcal{S}_n$ is the basic set of weighting functions, $p = |\mathcal{S}_n|$ satisfies $\log p \leqslant C_p \log n$ for a universal constant C_p, and so the second part of Assumption H3 holds for given κ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$. Furthermore, to prove the first part of Assumption H3, note that by Assumption L2, the event $\mathcal{A}_n$ that $h_{\max} = \max_{1 \leqslant i,j \leqslant n} |X_i - X_j|/2 \geqslant (C_2 - c_2)/4$ satisfies $\mathrm{P}_M(\mathcal{A}_n) = 1 - o(1)$ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$. Also, on $\mathcal{A}_n$, $h_{\min} = C_h h_{\max} (\log n/n)^{1/3} \geqslant c_h/n^{1/3}$ for some constant $c_h > 0$. Next, note that for any $s = (x, h) \in \mathcal{S}_n$, since

$$h \leqslant h_{\max} = \max_{1 \leqslant i,j \leqslant n} |X_i - X_j|/2 \leqslant (C_2 - c_2)/2,$$

either $c_2 + h \leqslant x$ or $x + h \leqslant C_2$ holds. Let $\mathcal{S}_{n,1}$ and $\mathcal{S}_{n,2}$ denote the subsets of those elements of $\mathcal{S}_n$ that satisfy the former and the latter inequalities, respectively, so that $\mathcal{S}_n = \mathcal{S}_{n,1} \cup \mathcal{S}_{n,2}$. Furthermore, let $C_K \in (0, 1)$ be some constant. Since the kernel function $K(\cdot)$ is continuous and strictly positive on the interior of its support, $\min_{t \in [-C_K, 0]} K(t) > 0$. In addition, since $K(\cdot)$ is continuous and has a bounded support, $K(\cdot)$ is bounded, and so it is possible to find a constant $c_K \in (0, 1)$ such that and $c_K + C_K \leqslant 1$ and

$$6c_K^{k+1} C_3 \max_{t \in [-1, -1+c_K]} K(t) \leqslant c_3 (1 - c_K - C_K)^k C_K \min_{t \in [-C_K, 0]} K(t), \tag{B.1}$$

where the constant k appears in the definition of the kernel weighting functions in Section 2.2.

Now, denote

$$\begin{aligned}
M_{n,1}(x,h) &= \Big\{ i = 1, \ldots, n \colon X_i \in [x - C_K h, x] \Big\}, \\
M_{n,2}(x,h) &= \Big\{ i = 1, \ldots, n \colon X_i \in [x - h, x - (1 - c_K)h] \Big\}, \\
M_{n,3}(x,h) &= \Big\{ i = 1, \ldots, n \colon X_i \in [x - (1 - c_K/2)h, x - (1 - c_K)h] \Big\}, \\
M_{n,4}(x,h) &= \Big\{ i = 1, \ldots, n \colon X_i \in [x - h, x + h] \Big\},
\end{aligned}$$

and let $\mathcal{B}_n$ be the event that

$$\begin{aligned}
(1/2) c_3 C_K nh \leqslant |M_{n,1}(x,h)| &\leqslant (3/2) C_3 C_K nh, \quad \text{for all } (x,h) \in \mathcal{S}_{n,1}, \\
(1/2) c_3 c_K nh \leqslant |M_{n,2}(x,h)| &\leqslant (3/2) C_3 c_K nh, \quad \text{for all } (x,h) \in \mathcal{S}_{n,1}, \\
(1/2) c_3 (c_K/2) nh \leqslant |M_{n,3}(x,h)| &\leqslant (3/2) C_3 (c_K/2) nh, \quad \text{for all } (x,h) \in \mathcal{S}_{n,1}, \\
(1/2) c_3 nh \leqslant |M_{n,4}(x,h)| &\leqslant (3/2) C_3 2nh, \quad \text{for all } (x,h) \in \mathcal{S}_n.
\end{aligned} \tag{B.2}$$

By Assumption 2 and Lemma C.1 in Appendix A.2, $\mathrm{P}_M(\mathcal{B}_n) = 1 - o(1)$ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$ since $h_{\min} \geqslant c_h/n^{1/3}$ on $\mathcal{A}_n$ and $\mathrm{P}_M(\mathcal{A}_n) = 1 - o(1)$ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$. In addition, for the constant

$$c = (1 - c_K - C_K)^k c_3 C_K \min_{t \in [-C_K, 0]} K(t)/4$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

and for all $s = (x,h) \in \mathcal{S}_{n,1}$ and $i \in M_{n,3}(x,h)$, on $\mathcal{B}_n$,

$$\begin{aligned}
&\sum_{1 \leqslant j \leqslant n} \operatorname{sign}(X_j - X_i)|X_j - X_i|^k K\Big(\frac{X_j - x}{h}\Big) \\
&\quad \geqslant \sum_{j \in M_{n,1}(x,h)} (1 - c_K - C_K)^k h^k K\Big(\frac{X_j - x}{h}\Big) - \sum_{j \in M_{n,2}(x,h)} c_K^k h^k K\Big(\frac{X_j - x}{h}\Big) \\
&\quad \geqslant (1 - c_K - C_K)^k h^k (1/2) c_3 C_K nh \min_{t \in [-C_K, 0]} K(t) - c_K^k h^k (3/2) C_3 c_K nh \max_{t \in [-1, -1 + c_k]} K(t) \\
&\quad \geqslant (1 - c_K - C_K)^k h^k c_3 C_K nh \min_{t \in [-C_K, 0]} K(t)/4 = cnh^{k+1},
\end{aligned}$$

where the inequality preceding the last one follows from (B.1). Hence, since for all $s = (x,h) \in \mathcal{S}_{n,1}$,

$$\begin{aligned}
V(s) &= \sum_{1 \leqslant i \leqslant n} \sigma_i^2 \left(\sum_{1 \leqslant j \leqslant n} \operatorname{sign}(X_j - X_i) Q(X_i, X_j, s) \right)^2 \\
&= \sum_{1 \leqslant i \leqslant n} \sigma_i^2 K\Big(\frac{X_i - x}{h}\Big)^2 \left(\sum_{1 \leqslant j \leqslant n} \operatorname{sign}(X_j - X_i)|X_j - X_i|^k K\Big(\frac{X_j - x}{h}\Big) \right)^2 \\
&\geqslant \sum_{i \in M_{n,3}(x,h)} \sigma_i^2 K\Big(\frac{X_i - x}{h}\Big)^2 \left(\sum_{1 \leqslant j \leqslant n} \operatorname{sign}(X_j - X_i)|X_j - X_i|^k K\Big(\frac{X_j - x}{h}\Big) \right)^2,
\end{aligned}$$

it follows that there exists a constant $c_{V,1} > 0$ such that on $\mathcal{B}_n$, $V(s) \geqslant c_{V,1}(nh)^3 h^{2k}$ for all $s = (x,h) \in \mathcal{S}_n$, and similar arguments also show that there exists a constant $c_{V,2} > 0$ such that the event $\mathcal{C}_n$ that $V(s) \geqslant c_{V,2}(nh)^3 h^{2k}$ for all $s = (x,h) \in \mathcal{S}_{n,2}$ satisfies $\mathrm{P}_M(\mathcal{C}_n) = 1 - o(1)$ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$. Then for the constant $c_V = \min(c_{V,1}, c_{V,2})$, on $\mathcal{B}_n \cap \mathcal{C}_n$,

$$V(s) \geqslant c_V (nh)^3 h^{2k}, \quad \text{for all } s = (x,h) \in \mathcal{S}_n. \tag{B.3}$$

Moreover, for the constant

$$C = 3 \cdot 2^k \cdot C_3 \left(\max_{t \in [-1,+1]} K(t) \right)^2,$$

on $\mathcal{B}_n$, by (B.2),

$$\left| \sum_{1 \leqslant j \leqslant n} \operatorname{sign}(X_j - X_i) Q(X_i, X_j, s) \right| \leqslant (2h)^k |M_{n,4}(x,h)| \left(\max_{t \in [-1,+1]} K(t) \right)^2 = Cnh^{k+1} \tag{B.4}$$

for all $i = 1, \ldots, n$ and $s = (x,h) \in \mathcal{S}_n$. Combining (B.4) with (B.3) shows that there exists a constant $C_A > 0$ such that on $\mathcal{B}_n \cap \mathcal{C}_n$,

$$A_n = \max_{s \in \mathcal{S}_n} \max_{1 \leqslant i \leqslant n} \left| \sum_{1 \leqslant j \leqslant n} \frac{\operatorname{sign}(X_j - X_i) Q(X_i, X_j, s)}{(V(s))^{1/2}} \right| \leqslant \frac{C_A}{(nh_{\min})^{1/2}},$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

and so

$$nA_n^4 \log^7(pn) \leqslant \frac{C_A^4 (C_p+1)^7 \log^7 n}{nh_{\min}^2} \leqslant \frac{C_A^4 (C_p+1)^7 \log^7 n}{c_h^2 n^{1/3}} = o(1).$$

Hence, given that $\mathrm{P}_M(\mathcal{B}_n \cap \mathcal{C}_n) = 1 - o(1)$ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$, it follows that the first part of Assumption H3 holds uniformly over $M \in \mathcal{M}_{\mathcal{L}}$.

Second, I verify Assumption H2. To do so, note that for the constant

$$C_D = 3^3 \cdot 2^{2k} \cdot C_3^3 \left(\max_{t \in [-1,+1]} K(t) \right)^4$$

and all $s \in (x,h) \in \mathcal{S}_n$, on $\mathcal{B}_n$, by (B.2),

$$\sum_{1 \leqslant i \leqslant n} \left(\sum_{1 \leqslant j \leqslant n} \operatorname{sign}(X_j - X_i) Q(X_i, X_j, s) \right)^2 \leqslant (2h)^{2k} |M_{n,4}(x,h)|^3 \left(\max_{t \in [-1,+1]} K(t) \right)^4 = C_D (nh)^3 h^{2k},$$

and so

$$|\widehat{V}(s) - V(s)| \leqslant \max_{1 \leqslant i \leqslant n} |\widehat{\sigma}_i^2 - \sigma_i^2| \times \sum_{1 \leqslant i \leqslant n} \left(\sum_{1 \leqslant j \leqslant n} \operatorname{sign}(X_j - X_i) Q(X_i, X_j, s) \right)^2 \leqslant C_D (nh)^3 h^{2k} \times \max_{1 \leqslant i \leqslant n} |\widehat{\sigma}_i^2 - \sigma_i^2|.$$

Combining this bound with (B.3) shows that on $\mathcal{B}_n \cap \mathcal{C}_n$,

$$|\widehat{V}(s)/V(s) - 1| = \frac{|\widehat{V}(s) - V(s)|}{V(s)} \leqslant (C_D/c_V) \max_{1 \leqslant i \leqslant n} |\widehat{\sigma}_i^2 - \sigma_i^2|$$

for all $s = (x,h) \in \mathcal{S}_n$. Hence, given that $\max_{1 \leqslant i \leqslant n} |\widehat{\sigma}_i^2 - \sigma_i^2| = o_p(n^{-\kappa})$ and $P_M(\mathcal{B}_n \cap \mathcal{C}_n) = 1 - o(1)$ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$, it follows that Assumption H2 holds uniformly over $M \in \mathcal{M}_{\mathcal{L}}$.

Finally, I verify Assumption H4. Recall that $h_n = (\log p/n)^{1/(2\beta+3)}$ and that $\log p \leqslant C_p \log n$. Also recall that on $\mathcal{A}_n$, $h_{\max} \leqslant (C_2 - c_2)/4$ and $h_{\min} \geqslant c_h/n^{1/3}$. Hence, the event $\mathcal{D}_n$ that there exists $\tilde{h} \in H_n$ such that $\tilde{h} \in (h_n/6, h_n/3]$ satisfies $\mathrm{P}_M(\mathcal{D}_n) = 1 - o(1)$ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$ since $\mathrm{P}_M(\mathcal{A}_n) = 1 - o(1)$ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$. In addition, the event $\mathcal{E}_n$ that for any $x_1, x_2 \in \mathcal{X}$ such that $x_2 - x_1 \geqslant h_n$, there exists $i_{x_1,x_2} = 1, \ldots, n$ such that $X_i \in [x_1 + h_n/3, x_2 - h_n/3]$ satisfies $\mathrm{P}_M(\mathcal{E}_n) = 1 - o(1)$ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$. But on $\mathcal{D}_n \cap \mathcal{E}_n$, the pair $(i_{x_1,x_2}, \tilde{h})$ satisfies conditions (i)–(iii) of Assumption H4. Hence, given that $\mathrm{P}_M(\mathcal{D}_n \cap \mathcal{E}_n) = 1 - o(1)$ uniformly over $M \in \mathcal{M}_{\mathcal{L}}$, it follows that Assumption H4 holds uniformly over $M \in \mathcal{M}_{\mathcal{L}}$. This completes the proof of the theorem. ■

Proof of Corollary 4.1. By Theorems 4.1 and 4.2, Assumptions H1–H4 hold uniformly over models $M \in \mathcal{M}_{\mathcal{L}}$. Hence, all the results of Theorems 3.1–3.4 hold with $\mathcal{M}_{\mathcal{H}} = \mathcal{M}_{\mathcal{L}}$. Moreover, $\log p$ in those results can be replaced by $\log n$ since, as in the proof of Theorem

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

4.2, $\log p \leqslant C_p \log n$ for some constant $C_p > 0$. This gives all the asserted claims of the corollary. ■

APPENDIX C: Useful Lemmas

LEMMA C.1. *Let $W_1, \ldots, W_n$ be an i.i.d. sequence of random variables with the support $[s_l, s_r]$ such that $c_1(x_2 - x_1) \leqslant \mathrm{P}(W_1 \in [x_1, x_2]) \leqslant C_1(x_2 - x_1)$ for some $c_1, C_1 > 0$ and all $[x_1, x_2] \subset [s_l, s_r]$. Then for any $c_2 > 0$, with probability $1 - o(1)$,*

$$(c_1/2)n(x_2 - x_1) \leqslant \left|\{i = 1, \ldots, n\colon W_i \in [x_1, x_2]\}\right| \leqslant (3C_1/2)n(x_2 - x_1) \tag{C.1}$$

simultaneously for all intervals $[x_1, x_2] \subset [s_l, s_r]$ satisfying $x_2 - x_1 \geqslant c_2(\log n)^2/n$. Moreover, the result holds uniformly over distributions of W_i's satisfying assumptions of the lemma with the same constants s_l, s_r, c_1, and C_1.

Proof. Let $K_n = [2(s_r - s_l)/(c_2(\log n)^2/n)]$ where $[x]$ denotes the largest integer smaller than or equal to x. Denote $\Delta_n = (s_r - s_l)/K_n$. For $k = 0, \ldots, K_m$, denote $y_{n,k} = s_l + k\Delta_n$. It suffices to show that (C.1) holds simultaneously for all intervals $[x_1, x_2]$ of the form $[y_{n,k-1}, y_{n,k}]$ for $k = 1, \ldots, K_n$.

Let $I_{i,k,n} = 1\{W_i \in [y_{n,k-1}, y_{n,k}]\}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K_n$. Then

$$\sum_{1 \leqslant i \leqslant n} I_{i,k,n} = \left|\{i = 1, \ldots, n\colon W_i \in [y_{n,k-1}, y_{n,k}]\}\right|, \quad k = 1, \ldots, K_n.$$

In addition,

$$\mathrm{E}[I_{i,k,n}] = \mathrm{E}[I_{i,k,n}^2] = \mathrm{P}(W_i \in [y_{n,k-1}, y_{n,k}]), \quad i = 1, \ldots, n \text{ and } k = 1, \ldots, K_n,$$

so that

$$c_1 n \Delta_n \leqslant \mathrm{E}\left[\sum_{1 \leqslant i \leqslant n} I_{i,k,n}\right] \leqslant C_1 n \Delta_n, \quad k = 1, \ldots, K_n,$$

and

$$\mathrm{Var}\left(\sum_{1 \leqslant i \leqslant n} I_{i,k,n}\right) \leqslant n\mathrm{E}[I_{i,k,n}^2] \leqslant C_1 n \Delta_n, \quad k = 1, \ldots, K_n.$$

Hence, by Bernstein's inequality (see Lemma 2.2.9 in van der Vaart and Wellner, 1996),

$$\mathrm{P}\left(\sum_{1 \leqslant i \leqslant n} I_{i,k,n} > (3/2)C_1 n \Delta_n\right) \leqslant \exp\left(-C(\log n)^2\right),$$

$$\mathrm{P}\left(\sum_{1 \leqslant i \leqslant n} I_{i,k,n} < (1/2)c_1 n \Delta_n\right) \leqslant \exp\left(-C(\log n)^2\right)$$

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

for some constant $C > 0$ that depends only on s_l, s_r, c_1, C_1, and c_2. Therefore, by the union bound,

$$\mathrm{P}\left(\sum_{1\leqslant i\leqslant n} I_{i,k,n} > (3/2)C_1 n\Delta_n \text{ or } \sum_{1\leqslant i\leqslant n} I_{i,k,n} < (1/2)c_1 n\Delta_n \text{ for some } k = 1,\dots,K_n\right)$$
$$\leqslant 2K_n \exp\left(-C(\log n)^2\right) = o(1),$$

where the last conclusion follows from the fact that $K_n \leqslant Cn$ for some constant $C > 0$. This gives the first asserted claim. The second asserted claim follows by noting that both the sequence $\{K_n\}_{n\geqslant 1}$ and the constant C depend only on s_l, s_r, c_1, C_1, and c_2. ■

LEMMA C.2. *Let $W_1,\dots,W_n$ be independent random vectors in $\mathbb{R}^p$ with $p \geqslant 2$. Let W_{ij} denote the jth component of W_i, that is $W_i = (W_{i1},\dots,W_{ip})^T$. Define $M = \max_{1\leqslant i\leqslant n}\max_{1\leqslant j\leqslant p}|W_{ij}|$ and $\sigma^2 = \max_{1\leqslant j\leqslant p}\sum_{1\leqslant i\leqslant n}\mathrm{E}[W_{ij}^2]$. Then*

$$\mathrm{E}\left[\max_{1\leqslant j\leqslant p}\left|\sum_{1\leqslant i\leqslant n}(W_{ij} - \mathrm{E}[W_{ij}])\right|\right] \leqslant C\left(\sigma\sqrt{\log p} + \sqrt{\mathrm{E}[M^2]}\log p\right)$$

for some universal $C > 0$.

Proof. See Lemma 8 in Chernozhukov et al. (2015). ■

LEMMA C.3. *Let $x_1,\dots,x_n$ be a sequence of independent zero-mean vectors in $\mathbb{R}^p$ with x_{ij} denoting the jth component of x_i, that is $x_i = (x_{i1},\dots,x_{ip})^T$. Let $y_1,\dots,y_n$ be a sequence of independent zero-mean Gaussian vectors in $\mathbb{R}^p$ with y_{ij} denoting the jth component of y_i, that is $y_i = (y_{i1},\dots,y_{ip})^T$. Assume that $\mathrm{E}[x_i x_i^T] = \mathrm{E}[y_i y_i^T]$ for all $i = \overline{1,n}$. Furthermore, assume that for all i and j, $x_{ij} = z_{ij}u_i$ where z_{ij}'s are nonstochastic with $|z_{ij}| \leqslant B_n$ and $\sum_{1\leqslant i\leqslant n} z_{ij}^2/n = 1$ where $\{B_n\}$ is a sequence of positive constants. Finally, assume that for some constants $c_1, C_1, c_2, C_2 > 0$ the following conditions hold: $\mathrm{E}[u_i^2] \geqslant c_1$, $\mathrm{E}[u_i^4] \leqslant C_1$, and $B_n^4\log^7(pn)/n \leqslant C_2 n^{-c_2}$. Then there exist constants $c, C > 0$ depending only on c_1, C_1, c_2, C_2 such that*

$$\sup_{t\in\mathbb{R}}\left|\mathrm{P}\left(\max_{1\leqslant j\leqslant p}\frac{1}{\sqrt{n}}\sum_{1\leqslant i\leqslant n} x_{ij} \leqslant t\right) - \mathrm{P}\left(\max_{1\leqslant j\leqslant p}\frac{1}{\sqrt{n}}\sum_{1\leqslant i\leqslant n} y_{ij} \leqslant t\right)\right| \leqslant Cn^{-c} \qquad \textbf{(C.2)}$$

for all n. In addition, if the terms $C_2 n^{-c_2}$ above are replaced by η_n where $\{\eta_n\}$ is a sequence of positive numbers converging to zero, then there exists another sequence $\{\eta_n'\}$ of positive numbers converging to zero and depending only on $\{\eta_n\}$ such that

$$\sup_{t\in\mathbb{R}}\left|\mathrm{P}\left(\max_{1\leqslant j\leqslant p}\frac{1}{\sqrt{n}}\sum_{1\leqslant i\leqslant n} x_{ij} \leqslant t\right) - \mathrm{P}\left(\max_{1\leqslant j\leqslant p}\frac{1}{\sqrt{n}}\sum_{1\leqslant i\leqslant n} y_{ij} \leqslant t\right)\right| \leqslant \eta_n' \qquad \textbf{(C.3)}$$

for all n.

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press

Proof. The result in (C.2) is proven in Corollary 2.1 of Chernozhukov et al. (2013). Furthermore, inspecting the proof of Corollary 2.1 of Chernozhukov et al. (2013) shows that the sequences $C_2 n^{-c_2}$ and Cn^{-c} in (C.2) can be replaced by general sequences $\{\eta_n\}$ and $\{\eta_n'\}$ of positive numbers converging to zero, and so the result in (C.3) holds as well. ■

https://doi.org/10.1017/S0266466618000282 Published online by Cambridge University Press