Chetverikov, Denis

Working Paper

Testing regression monotonicity in econometric models

cemmap working paper, No. CWP35/12

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Chetverikov, Denis (2012) : Testing regression monotonicity in econometric models, cemmap working paper, No. CWP35/12, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2012.3512

This Version is available at:
https://hdl.handle.net/10419/79545

cemmap
centre for microdata methods and practice

Testing regression monotonicity in econometric models

Denis Chetverikov

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP35/12

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

TESTING REGRESSION MONOTONICITY IN ECONOMETRIC MODELS

DENIS CHETVERIKOV

Abstract. 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 paper 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 paper 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 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). An application of the developed procedures to the dataset of Ellison and Ellison (2011) shows that there is some evidence of strategic entry deterrence in pharmaceutical industry where incumbents may use strategic investment to prevent generic entries when their patents expire.

1. Introduction

The concept of monotonicity often appears in economics research. For example, monotone comparative statics has been a popular research topic in economic theory for many years. See, in particular, the seminal work on this topic by Milgrom and Shannon (1994) and Athey (2002). Given the great deal of effort put into deriving conditions that are necessary and sufficient for monotonicity in theoretical models, the natural question is whether we observe monotonicity in the data. This paper provides a general nonparametric framework for testing monotonicity of a regression function. Tests of monotonicity developed in this paper can be used to evaluate assumptions and implications of economic theory concerning monotonicity. In addition, as was recently noticed by Ellison and Ellison (2011), these tests can also be used to provide evidence of

Date: First version: March 2012. This version: November 7, 2012. Email: dchetver@mit.edu. I thank Victor Chernozhukov for encouragement and guidance. I am also grateful to Anna Mikusheva, Isaiah Andrews, Glenn Ellison, Jose Montiel, and Whitney Newey for valuable comments. The first version of the paper was presented at the Econometrics lunch at MIT in April, 2012.

existence of certain phenomena related to strategic behavior of economic agents that are difficult to detect otherwise. Several motivating examples are presented in the next section.

I start with the model

$$Y_i = f(X_i) + \varepsilon_i,\ i = 1, 2, 3, ... \tag{1}$$

where Y_i is a scalar random variable, $\{X_i\} \subset \mathbb{R}$ is a sequence of nonstochastic design points, f is an unknown function, and $\{\varepsilon_i\}$ is a sequence of independent zero-mean unobserved scalar random variables. Later on in the paper, I extend the analysis to cover models with multivariate X_i's. I am interested in testing the null hypothesis, $\mathcal{H}_0$, that $f(x)$ 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 the sample of size n, $\{X_i, Y_i\}_{1 \leqslant i \leqslant n}$. I assume that f 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 functions with Lipschitz first order derivatives. Moreover, the rate optimal test constructed in this paper is adaptive in the sense that it automatically adapts to the unknown smoothness of f.

This paper makes several contributions. First, I introduce a general framework for testing monotonicity. This framework allows me to develop a broad class of new tests, which also includes some existing tests as special cases. This gives a researcher a lot of flexibility to incorporate ex ante information she might have. Second, I develop new methods to simulate the critical values for these tests that in many cases yield higher power than that of existing methods. Third, I consider the problem of testing for monotonicity in models with multiple covariates for the first time in the literature. As will be explained in the paper, these models are more difficult to analyze and require rather different treatment in comparison with the case of univariate X_i's.

Constructing a critical value is an important and difficult problem in nonparametric testing. The problem arises because most test statistics studied in the literature have some asymptotic distribution when f is constant but diverge if f is strictly increasing. This discontinuity implies that for some sequences of models $f = f_n$, the limit distribution depends on the local slope function, which is an unknown infinite-dimensional nuisance parameter that can not be estimated consistently from the data. A common approach in the literature to solve this problem is to calibrate the critical value using the case when the type I error is maximized (the least favorable model), i.e. the model with constant f.[1] In contrast, I develop two selection procedures that estimate the set where f 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 valuable power improvements in

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

comparison with other tests if f is strictly increasing over some subsets of its domain. The first selection procedure, which is based on the one-step approach, is related to those developed in Chernozhukov, Lee, and Rosen (2009), Andrews and Shi (2010), and Chetverikov (2012), all of which deal with the problem of testing conditional moment inequalities. The second selection procedure is based on the stepdown approach. It is related to methods developed in Romano and Wolf (2005b) and Romano and Shaikh (2010). The details, however, are rather different.

Another important issue in nonparametric testing is how to choose a smoothing parameter. In theory, the optimal smoothing parameter can be derived for many smoothness classes of functions f. In practice, however, the smoothness class that f belongs to is usually unknown. I deal with this problem by employing the adaptive testing approach. This allows me to obtain tests with good power properties when the information about smoothness of the function f 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 a 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. The tests of Gijbels, Hall, Jones, and Koch (2000) and Ghosal, Sen, and van der Vaart (2000) (from now on, GHJK and GSV, respectively) are based on the signs of $(Y_{i+k} - Y_i)(X_{i+k} - X_i)$. Hall and Heckman (2000) (from now on, HH) developed a test based on the slopes of local linear estimates of f. The list of other papers includes Schlee (1982), Bowman, Jones, and Gijbels (1998), Dumbgen and Spokoiny (2001), Durot (2003), Beraud, Huet, and Laurent (2005), and Wang and Meyer (2011). Lee, Linton, and Whang (2009) and Delgado and Escanciano (2010) derived tests of stochastic monotonicity, which means that the conditional cdf of Y given X, $F_{Y|X}(y, x)$, is (weakly) decreasing in x for any fixed y.

As an empirical application of the results developed in this paper, I consider the problem of detecting strategic entry deterrence in the pharmaceutical industry. In that industry, incumbents whose drug patents are about to expire can change their investment behavior in order to prevent generic entries after the expiration of the patent. Although there are many theoretically compelling arguments as to how and why incumbents should change their investment behavior (see, for example, Tirole (1988)), the empirical evidence is rather limited. Ellison and Ellison (2011) showed that, under certain conditions, the dependence of investment on market size should be monotone if no strategic entry deterrence is present. In addition, they noted that the entry deterrence motive should be important in intermediate-sized markets and less important in small and large markets. Therefore, strategic entry deterrence might result in the nonmonotonicity of

the relation between market size and investment. Hence, rejecting the null hypothesis of monotonicity provides the evidence in favor of the existence of strategic entry deterrence. I apply the tests developed in this paper to Ellison and Ellison's dataset and show that there is some evidence of nonmonotonicity in the data. The evidence is rather weak, though.

The rest of the paper is organized as follows. Section 2 provides motivating examples. Section 3 describes the general test statistic and gives several methods to simulate the critical value. Section 4 contains the main results under high-level conditions. Section 5 is devoted to the verification of high-level conditions under primitive assumptions. Since in most practically relevant cases, the model also contains some additional covariates, Section 6 studies the cases of partially linear and fully nonparametric models with multiple covariates. Section 7 presents a small Monte Carlo simulation study. Section 8 describes the empirical application. Section 9 concludes. All proofs are contained in the Appendix.

Notation. Throughout this paper, let $\{\epsilon_i\}$ denote a sequence of independent $N(0,1)$ random variables that are independent of the data. The sequence $\{\epsilon_i\}$ will be used in bootstraping critical values. The notation $i = \overline{1,n}$ is shorthand for $i \in \{1,...,n\}$. For any set $\mathcal{S}$, I denote the number of elements in this set by $|\mathcal{S}|$. The notation $a_n \lesssim b_n$ means that there exists a constant C independent of n such that $a_n \leqslant Cb_n$. I use symbol C to denote a generic constant the value of which may vary from line to line, and I use symbol C_j for an integer j to denote a constant the value of which is fixed throughout the paper.

2. Motivating Examples

There are many interesting examples where testing for monotonicity can be fruitfully used in economics. Several examples are provided in this section.

1. Testing implications of economic theory. Many testable implications of economic theory are concerned with comparative statics analysis. These implications most often take the form of qualitative statements like "Increasing factor X will positively (negatively) affect response variable Y". The common approach to test such results on the data is to look at the corresponding coefficient in the linear (or other parametric) regression. It is said that the theory is confirmed if the coefficient is significant and has the expected sign. More precisely, one should say that the theory is "confirmed on average" because the linear regression gives average coefficients. This approach can be complemented by testing monotonicity. If the hypothesis of monotonicity is rejected, it means that the theory is lacking some empirically important features.

For example, a classical paper Holmstrom and Milgrom (1994) on the theory of the firm is built around the observation that in multitask problems different incentive instruments are expected to be complementary to each other. Indeed, increasing an incentive for one task may lead the agent to spend too much time on that task ignoring other responsibilities. This can be avoided

if incentives on different tasks are balanced with each other. To derive testable implications of the theory, Holmstrom and Milgrom study a model of industrial selling introduced in Anderson and Schmittlein (1984) where a firm chooses between an in-house agent and an independent representative who divide their time into four tasks: (i) direct sales, (ii) investing in future sales to customers, (iii) nonsale activities, such as helping other agents, and (iv) selling the products of other manufacturers. Proposition 4 in their paper states that under certain conditions, the conditional probability of having an in-house agent is a (weakly) increasing function of the marginal cost of evaluating performance and is a (weakly) increasing function of the importance of nonselling activities. These are hypotheses that can be directly tested on the data by procedures developed in this paper. This would be an important extension of linear regression analysis performed, for example, in Anderson and Schmittlein (1984) and Poppo and Zenger (1998).

2. Testing assumptions of economic theory. Monotonicity is also a key assumption in many economic models, especially in those concerning equilibrium analysis. For example, in the theory of global games it is often assumed that the profit function of an individual given that she chooses a particular action is nondecreasing in the proportion of her opponents who also choose this action, or/and that this function is nondecreasing in an exogenous parameter. See, for example, Morris and Shin (1998), Morris and Shin (2001), and Angeletos and Werning (2006).

3. Detecting strategic effects. Certain strategic effects, the existence of which is difficult to prove otherwise, can be detected by testing for monotonicity. An example on strategic entry deterrence in the pharmaceutical industry is described in the Introduction and is analyzed in Section 8. Below I provide another example concerned with the problem of debt pricing. Consider a model where investors hold a collateralized debt. The debt will yield a fixed payment in the future if it is rolled over and an underlying project is successful. Otherwise the debt will yield nothing. Alternatively, all investors have an option of not rolling over and getting the value of the collateral immediately. The probability that the project turns out to be successful depends on the fundamentals and on how many investors roll over. Each investor possesses some information on the fundamentals. If this information is common knowledge, the price of the debt is clearly an increasing function of the value of the collateral. Morris and Shin (2004) show, however, that in the absence of common knowledge, high value of the collateral leads investors to believe that many other investors will not roll over, and the project will not be successful. This strategic effect implies that the price of the debt might decrease as the collateral becomes more valuable, thus causing nonmonotonicity. They argue that this effect is important for understanding anomalies in empirical implementation of the standard debt pricing theory of Merton (1974). A natural question is how to prove existence of this effect in the data. One possible strategy is to test whether conditional mean of the price of the debt given the value of the

collateral is a monotonically increasing function. Rejecting the null hypothesis of monotonicity provides evidence in favor of the existence of the strategic effect.

4. Testing assumptions of econometric models. Monotonicity is often assumed in the econometrics literature on estimating treatment effects. A widely used econometric model in this literature is as follows. Suppose that we observe a sample of individuals, $i = \overline{1,n}$. Each individual has a random response function $y_i(t)$ that gives her response for each level of treatment $t \in T$. Let z_i and $y_i = y_i(z_i)$ denote the realized level of the treatment and the realized response correspondingly (both of them are observable). The problem is how to derive inference on $E[y_i(t)]$. Manski and Pepper (2000) introduced assumptions of monotone treatment response, which imposes that $y_i(t_2) \geqslant y_i(t_1)$ whenever $t_2 \geqslant t_1$, and monotone treatment selection, which imposes that $E[y_i(t)|z_i = v]$ is increasing in v for all $t \in T$. The combination of these assumptions yields a testable prediction. Indeed, for all $v_2 \geqslant v_1$,

$$\begin{aligned} E[y_i|z_i = v_2] &= E[y_i(v_2)|z_i = v_2] \\ &\geqslant E[y_i(v_1)|z_i = v_2] \\ &\geqslant E[y_i(v_1)|z_i = v_1] \\ &= E[y_i|z_i = v_1]. \end{aligned}$$

Since all variables on both the left and right hand sides of this chain of inequalities are observable, this prediction can be tested by the procedures developed in this paper.

5. Classification problems. Some concepts in economics are defined using monotonicity. For example, a good is called normal (inferior) if demand for this good is an increasing (decreasing) function of income. A good is called luxury (necessity) if the share of income spent on this good is an increasing (decreasing) function of income. Monotonicity testing can be fruitfully used to classify different goods using this standard terminology. A related problem arises in the Ramsey-Cass-Koopman growth model where one of the most important questions is whether current savings is a nondecreasing function of current level of capital. See, for example, Milgrom and Shannon (1994).

3. The Test

3.1. **The General Test Statistic.** Recall that I consider a model given in equation (1), and the test should be based on the sample $\{X_i, Y_i\}_{i=1}^n$ of n observations where X_i and Y_i are a nonstochastic design point and a scalar dependent random variable, respectively. In this Section and in Sections 4 and 5, I assume that $X_i \in \mathbb{R}$. The case where $X_i \in \mathbb{R}^d$ for $d > 1$ is considered in Section 6.

Let $Q(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be some weighting function satisfying $Q(x_1, x_2) = Q(x_2, x_1)$ and $Q(x_1, x_2) \geqslant 0$ for all $x_1, x_2 \in \mathbb{R}$, and let

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

be a test function. Since $Q(X_i, X_j) \geqslant 0$ and $\mathrm{E}[Y_i] = f(X_i)$, it is easy to see that under $\mathcal{H}_0$, that is, when the function f is non-decreasing, $\mathrm{E}[b] \leqslant 0$. On the other hand, if $\mathcal{H}_0$ is violated, there exists a function $Q(\cdot,\cdot)$ such that $\mathrm{E}[b] > 0$. Therefore, b can be used to form a test statistic if I can find an appropriate function $Q(\cdot,\cdot)$. For this purpose, I will use the adaptive testing approach developed in statistics literature. Even though this approach has attractive features, it is almost never used in econometrics. An exception is Horowitz and Spokoiny (2001), who used it for specification testing.

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 for $s \in \mathcal{S}_n$, let $Q(\cdot,\cdot,s) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be some function satisfying $Q(x_1, x_2, s) = Q(x_2, x_1, s)$ and $Q(x_1, x_2, s) \geqslant 0$ for all $x_1, x_2 \in \mathbb{R}$. In addition, let

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

be a test function. Since X_i are nonstochastic, the variance of $b(s)$ is given by

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

where $\sigma_i = (\mathrm{E}[\varepsilon_i^2])^{1/2}$. In general, σ_i are unknown, and should be estimated from the data. Let $\widehat{\sigma}_i$ denote some (not necessarily consistent) estimator of σ_i. Available estimators are discussed later in this Section. Then the estimated variance of $b(s)$ is

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

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

$$T = T(\{X_i, Y_i\}, \{\widehat{\sigma}_i\}, \mathcal{S}_n) = \max_{s \in \mathcal{S}_n} \frac{b(\{X_i, Y_i\}, s)}{\sqrt{\widehat{V}(\{X_i\}, \{\widehat{\sigma}_i\}, s)}}.$$

Large values of T indicate that the null hypothesis is violated. Later on in this section, I will provide methods for estimating 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 types of deviations from $\mathcal{H}_0$. Indeed, each weighting function $Q(\cdot,\cdot,s)$ is useful for detecting a particular type of deviations, and so the larger the set of weighting functions

$\mathcal{S}_n$ is, the more types of deviations can be detected, and the higher is adaptivity of the test. In this paper, 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 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, in many cases the loss of power is relatively small; see, in particular, discussion after Theorem 2 on the dependence of critical values on the size of the set $\mathcal{S}_n$.

3.2. **Typical Weighting Functions.** Let me now describe typical weighting functions. Consider some positive compactly supported kernel function $K : \mathbb{R} \to \mathbb{R}$.[2] For convenience, I will assume that the support of K is $[-1, 1]$. In addition, let $s = (x, h)$ where x is a location point and h is a bandwidth value. Finally, define

$$Q(x_1, x_2, (x, h)) = |x_1 - x_2|^k K((x_1 - x)/h)K((x_2 - x)/h) \tag{2}$$

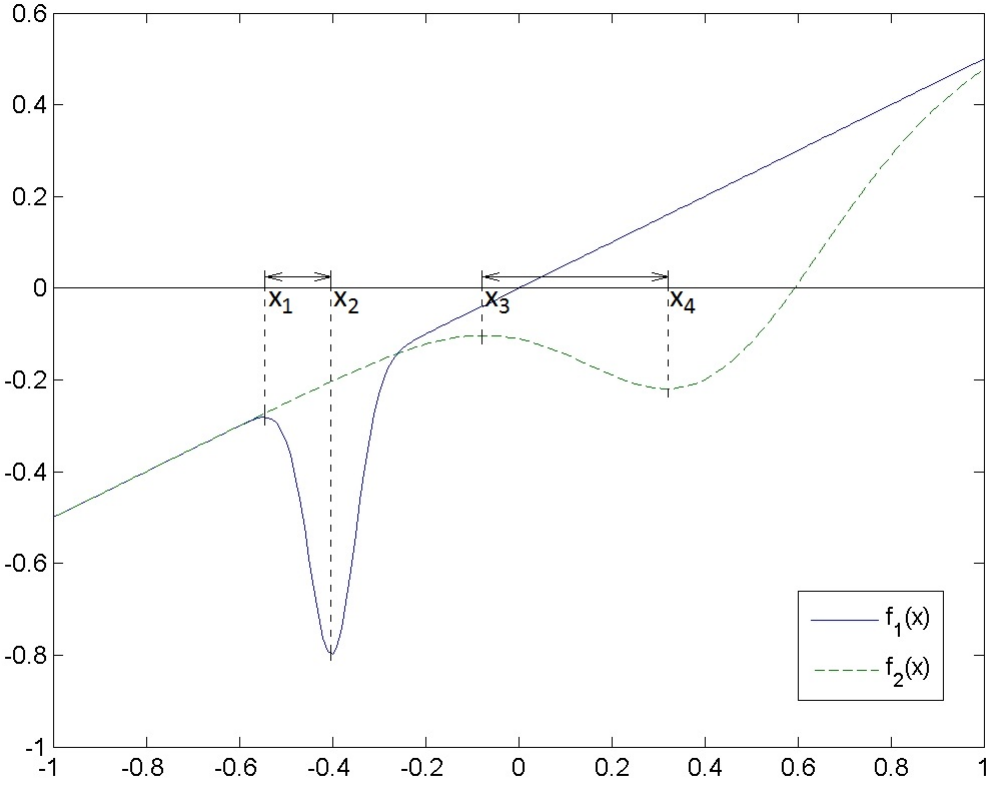
for some $k \geqslant 0$. I refer to this Q as a kernel weighting function.

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. To explain why this test has good adaptivity properties, consider figure 1 that plots two regression functions. Both f_1 and f_2 violate $\mathcal{H}_0$ but locations where $\mathcal{H}_0$ is violated are different. In particular, f_1 violates $\mathcal{H}_0$ on the interval $[x_1, x_2]$ while the corresponding interval for f_2 is $[x_3, x_4]$. In addition, f_1 is relatively less smooth than f_2, and $[x_1, x_2]$ is shorter than $[x_3, x_4]$. To have good power against f_1, $\mathcal{S}_n$ 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 of 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, $\mathcal{S}_n$ 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. Further, note that larger values of h yield smaller variance of b. So, given that $[x_3, x_4]$ is longer than $[x_1, x_2]$, the optimal pair (x, h) to test against f_2 has larger value of h than that to test against f_1. 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.

The general framework considered here gives the researcher a lot of flexibility in determining what weighting functions to use. In particular, if the researcher expects that any deviations from $\mathcal{H}_0$, if present, are concentrated around some particular point X_i, then she can restrict the set $\mathcal{S}_n$ and consider only pairs with $x = X_i$. Note that this will increase the power of the test because

[2]The kernel function is called positive if it is positive on its support.

FIGURE 1. Regression Functions Illustrating Different Deviations from $\mathcal{H}_0$



smaller sets $\mathcal{S}_n$ yield lower critical values. In addition, if it is expected that the function f is rather smooth, then the researcher 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.

Another interesting choice of the weighting functions is

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

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

If no ex ante information is available, I recommend using kernel weighting functions with $\mathcal{S}_n = \{(x, h) : x \in \{X_1, ..., X_n\}, h \in H_n\}$ where $H_n = \{h = h_{\max} u^l : h \geqslant h_{\min}, l = 0, 1, 2, ...\}$ and $h_{\max} = \max_{1 \leqslant i,j \leqslant n} |X_i - X_j|/2$. I refer to this $\mathcal{S}_n$ as a basic set of weighting functions. I also recommend setting $u = 0.5$, $h_{\min} = h_{\max}(0.3/n^{0.95})^{1/3}$, and $k = 0$ or 1. This choice of parameters is consistent with the theory presented in sections 4 and 5 and has worked well in simulations.

The value of $h_{\min}$ is selected so that the test function $b(s)$ for any given s uses no less than approximately 15 observations when $n = 100$.

3.3. **Comparison with Other Known Tests.** I will now show that the general framework described above includes the HH test statistic and a slightly modified version of the GSV test statistic as special cases that correspond to different values of k in the definition of kernel weighting functions.

GSV use the following test function:

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

whereas setting $k = 0$ in equation (2) yields

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

and so the only difference is that I include the term $(Y_i - Y_j)$ whereas they use $\text{sign}(Y_i - Y_j)$. It will be shown in the next section that my test is consistent. On the other hand, I claim that GSV test is not consistent under the presence of conditional heteroscedasticity. Indeed, assume 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 $f(X_i) = 0$ for $i = \overline{1, n}$. 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. Moreover, they consider a unique nonstochastic value of h, which means that the GSV test is nonadaptive with respect to the smoothness of the function f.

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

$$b(s) = \frac{\sum_{s_1 < i \leqslant s_2} Y_i \sum_{s_1 < j \leqslant s_2} (X_i - X_j)}{(s_2 - s_1) \sum_{s_1 < i \leqslant s_2} X_i^2 - (\sum_{s_1 < i \leqslant s_2} X_i)^2}, \tag{3}$$

where $s = (s_1, s_2)$. Note that the denominator of (3) is nonstochastic, and so it disappears after studentization. In addition, simple rearrangements show that the numerator in (3) is

$$(1/2) \sum_{1 \leqslant i,j \leqslant n} (Y_i - Y_j)(X_j - X_i)1\{x - h \leqslant X_i \leqslant x + h\}1\{x - h \leqslant X_j \leqslant x + h\} \tag{4}$$

for some x and h. On the other hand, setting $k = 1$ in equation (2) yields

$$b(s) = (1/2) \sum_{1 \leqslant i,j \leqslant n} (Y_i - Y_j)(X_j - X_i)K((X_i - x)/h)K((X_j - x)/h). \tag{5}$$

Noting that expression in (4) is proportional to that on the right hand side in (5) with $K(\cdot) = 1\{[-1,+1]\}(\cdot)$ implies that the HH test statistic is a special case of those studied in this paper.

3.4. **Estimating σ_i.** In practice, σ_i is usually unknown, and, hence, should be estimated from the data. Let $\widehat{\sigma}_i$ denote some estimator of σ_i. I provide results for two types of estimators. The first type of estimators is easier to implement but the second worked better in simulations.

First, σ_i can be estimated by the residual $\widehat{\varepsilon}_i$. More precisely, let $\widehat{f}$ be some uniformly consistent estimator of f with at least a polynomial rate of consistency in probability, i.e. $\widehat{f}(X_i) - f(X_i) = o_p(n^{-\kappa_1})$ uniformly over $i = \overline{1,n}$ for some $\kappa_1 > 0$, and let $\widehat{\sigma}_i = \widehat{\varepsilon}_i$ where $\widehat{\varepsilon}_i = Y_i - \widehat{f}(X_i)$. Note that $\widehat{\sigma}_i$ can be negative. Clearly, $\widehat{\sigma}_i$ is not a consistent estimator of σ_i. Nevertheless, as I will show in Section 4, this estimator leads to valid inference. Intuitively, it works because the test statistic contains the weighted average sum of σ_i^2, $i = \overline{1,n}$, and the estimation error averages out. To obtain a uniformly consistent estimator $\widehat{f}$ of f, one can use a series method (see Newey (1997), theorem 1) or local polynomial regression (see Tsybakov (2009), theorem 1.8). If one prefers kernel methods, it is important to use generalized kernels in order to deal with boundary effects when higher order kernels are used; see, for example, Muller (1991). Alternatively, one can choose $\mathcal{S}_n$ so that boundary points are excluded from the test statistic. In addition, if the researcher decides to impose some parametric structure on the set of potentially possible functions, then parametric methods like OLS will typically give uniform consistency with κ_1 arbitrarily close to $1/2$.

The second way of estimating σ_i is to use a parametric or nonparametric estimator $\widehat{\sigma}_i$ satisfying $\widehat{\sigma}_i - \sigma_i = o_p(n^{-\kappa_1})$ uniformly over $i = \overline{1,n}$ for some $\kappa_1 > 0$. Many estimators of σ_i satisfy this condition. Assume that the data $\{X_i, Y_i\}_{i=1}^n$ are arranged so that $X_i \leqslant X_j$ whenever $i \leqslant j$. Then the estimator of Rice (1984), given by

$$\widehat{\sigma} = \left(\frac{1}{2n} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2 \right)^{1/2}, \tag{6}$$

is $\sqrt{n}$-consistent if $\sigma_i = \sigma$ for all $i = \overline{1,n}$ and f is piecewise Lipschitz-continuous.

The Rice estimator can be easily modified to allow for conditional heteroscedasticity. Choose a bandwidth value $b_n > 0$. For $i = \overline{1,n}$, let $J(i) = \{j = \overline{1,n} : |X_j - X_i| \leqslant b_n\}$. Let $|J(i)|$ denote the number of elements in $J(i)$. Then σ_i can be estimated by

$$\widehat{\sigma}_i = \left(\frac{1}{2|J(i)|} \sum_{j \in J(i): j+1 \in J(i)} (Y_{j+1} - Y_j)^2 \right)^{1/2}. \tag{7}$$

I refer to (7) as a local version of Rice's estimator. An advantage of this estimator is that it is adaptive with respect to the smoothness of the function f. Lemma 2 in Section 5 provides conditions that are sufficient for uniform consistency of this estimator with at least a polynomial

rate. The key condition there is that $|\sigma_{j+1} - \sigma_j| \leqslant C|X_{j+1} - X_j|$ for some $C > 0$ and all $j = \overline{1, n-1}$. The intuition for consistency is as follows. Note that X_{j+1} is close to X_j. So, if the function f is continuous, then

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

so that

$$E[(Y_{j+1} - Y_j)^2] \approx \sigma_{j+1}^2 + \sigma_j^2$$

since ε_{j+1} is independent of ε_j. Further, if b_n is sufficiently small, then $\sigma_{j+1}^2 + \sigma_j^2 \approx 2\sigma_i^2$ since $|X_{j+1} - X_i| \leqslant b_n$ and $|X_j - X_i| \leqslant b_n$, and so $\widehat{\sigma}_i^2$ is close to σ_i. Other available estimators are presented, for example, in Muller and Stadtmuller (1987), Fan and Yao (1998), Horowitz and Spokoiny (2001), and Hardle and Tsybakov (1997).

3.5. **Simulating the Critical Value.** In this subsection, I provide three different methods for estimating quantiles of the null distribution of the test statistic T. These are plug-in, one-step, and stepdown 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 by Liu (1988), Mammen (1993), Hardle and Mammen (1993), Horowitz and Spokoiny (2001), and Chetverikov (2012). See also Chernozhukov, Chetverikov, and Kato (2012). The three methods are arranged in terms of increasing power and computational complexity. The validity of all three methods is established in theorem 1. Recall that $\{\epsilon_i\}$ denotes a sequence of independent $N(0,1)$ random variables that are independent of the data.

Plug-in Approach. Suppose that we want to obtain a test of size α. The plug-in approach is based on two observations. First, under $\mathcal{H}_0$,

$$\begin{aligned} b(s) &= (1/2) \sum_{1 \leqslant i,j \leqslant n} (f(X_i) - f(X_j) + \varepsilon_i - \varepsilon_j)\text{sign}(X_j - X_i)Q(X_i, X_j, s) \quad (8) \\ &\leqslant (1/2) \sum_{1 \leqslant i,j \leqslant n} (\varepsilon_i - \varepsilon_j)\text{sign}(X_j - X_i)Q(X_i, X_j, s) \quad (9) \end{aligned}$$

since $Q(X_i, X_j) \geqslant 0$ and $f(X_i) \geqslant f(X_j)$ whenever $X_i \geqslant X_j$ under $\mathcal{H}_0$, and so the $(1-\alpha)$ quantile of T is bounded from above by the $(1-\alpha)$ quantile of T in the model with $f(x) = 0$ for all $x \in \mathbb{R}$, which is the least favorable model under $\mathcal{H}_0$. Second, it will be shown that the distribution of T asymptotically depends on the distribution of noise $\{\varepsilon_i\}$ only through $\{\sigma_i^2\}$. These two observations suggest that the critical value for the test can be obtained by simulating the conditional $(1-\alpha)$ quantile of $T^\star = T(\{X_i, Y_i^\star\}, \{\widehat{\sigma}_i\}, \mathcal{S}_n)$ given $\{\widehat{\sigma}\}$ where $Y_i^\star = \widehat{\sigma}_i \epsilon_i$ for $i = \overline{1, n}$. This is called the plug-in critical value $c_{1-\alpha}^{PI}$. See section A of the Appendix for detailed step-by-step instructions.

One-Step Approach. The test with the plug-in critical value is computationally rather simple. It has, however, poor power properties. Indeed, the distribution of T in general depends on f but the plug-in approach is based on the least favorable regression function $f = 0$, and so it is too conservative when f is strictly increasing. More formally, suppose for example that a kernel weighting function is used, and that f is strictly increasing in h-neighborhood of X_i but is constant in h-neighborhood of X_j. Let $s_1 = s(X_i, h)$ and $s_2 = s(X_j, h)$. Then $b(s_1)/(\widehat{V}(s_1))^{1/2}$ is no greater than $b(s_2)/(\widehat{V}(s_2))^{1/2}$ with probability approaching one. On the other hand, $b(s_1)/(\widehat{V}(s_1))^{1/2}$ is greater than $b(s_2)/(\widehat{V}(s_2))^{1/2}$ with nontrivial probability in the model with $f(x) = 0$ for all $x \in \mathbb{R}$, which is used to obtain $c^{PI}_{1-\alpha}$. Therefore, $c^{PI}_{1-\alpha}$ overestimates the corresponding quantile of T. The natural idea to overcome the conservativeness of the plug-in approach is to simulate a critical value using not all elements of $\mathcal{S}_n$ but only those that are relevant for the given sample. Two selection procedures developed in this paper are used to decide what elements of $\mathcal{S}_n$ should be used in the simulation. The main difficulty here is to make sure that the selection procedures do not distort the size of the test. The simpler of these two procedures is the one-step approach.

Let $\{\gamma_n\}$ 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 addition, denote

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

Then the one-step critical value $c^{OS}_{1-\alpha}$ is the conditional $(1-\alpha)$ quantile of the simulated statistic $T^\star = T(\{X_i, Y_i\star\}, \{\widehat{\sigma}_i\}, \mathcal{S}^{OS}_n)$ given $\{\widehat{\sigma}_i\}$ and $\mathcal{S}^{OS}_n$ where $Y^\star_i = \widehat{\sigma}_i\epsilon_i$ for $i = \overline{1,n}$.[3] Intuitively, the one-step critical value works because the weighting functions corresponding to elements of the set $\mathcal{S}_n \backslash \mathcal{S}^{OS}_n$ have an asymptotically negligible influence on the distribution of T under $\mathcal{H}_0$. Indeed, the probability that at least one element s of $\mathcal{S}_n$ such that

$$(1/2) \sum_{1 \leqslant i,j \leqslant n} (f(X_i) - f(X_j))\text{sign}(X_j - X_i)Q(X_i, X_j, s)/(\widehat{V}(s))^{1/2} > -c^{PI}_{1-\gamma_n} \tag{10}$$

belongs to the set $\mathcal{S}_n \backslash \mathcal{S}^{OS}_n$ is at most $\gamma_n + o(1)$. On the other hand, the probability that at least one element s of $\mathcal{S}_n$ such that inequality (10) does not hold for this element gives $b(s)/(\widehat{V}(s))^{1/2} > 0$ is again at most $\gamma_n + o(1)$. Since γ_n converges to zero, this suggests that the critical value can be simulated using only elements of $\mathcal{S}^{OS}_n$. In practice, one can set γ_n as a small fraction of α. For example, the Monte Carlo simulations presented in this paper use $\gamma_n = 0.01$ with $\alpha = 0.1$.

Stepdown Approach. The one-step approach, as the name suggests, uses only one step to cut out 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 stepdown procedure and yielding second-order improvements in the power. The stepdown procedures were developed in the literature on multiple hypothesis testing; see, in particular, Holm (1979), Romano and Wolf (2005a), Romano and Wolf (2005b),

[3] As usual, I define the maximum over the empty set as $+\infty$, and so $c^{OS}_{1-\alpha} = +\infty$ if $\mathcal{S}^{OS}_n$ is empty.

and Romano and Shaikh (2010), and Lehmann and Romano (2005) for a textbook introduction. The use of stepdown method in this paper, however, is rather different.

To explain the stepdown approach, let me define the sequences $(c^l_{1-\gamma_n})_{l=1}^\infty$ and $(\mathcal{S}^l_n)_{l=1}^\infty$. 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 conditional $(1-\gamma_n)$ quantile of $T^\star = T(\{X_i, Y_i^\star\}, \{\widehat{\sigma}_i\}, \mathcal{S}^l_n)$ given $\{\widehat{\sigma}_i\}$ and $\mathcal{S}^l_n$ where $Y_i^\star = \widehat{\sigma}_i \epsilon_i$ for $i = \overline{1,n}$ and

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

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, $\mathcal{S}^{l(0)}_n = \mathcal{S}^{l(0)+1}_n$ for some $l(0) \geqslant 1$ and $\mathcal{S}^l_n = \mathcal{S}^{l+1}_n$ for all $l \geqslant l(0)$. Let $\mathcal{S}^{SD}_n = \mathcal{S}^{l(0)}_n$. Then the stepdown critical value $c^{SD}_{1-\alpha}$ is the conditional $(1-\alpha)$ quantile of $T^\star = T(\{X_i, Y_i^\star\}, \{\widehat{\sigma}_i\}, \mathcal{S}^{SD}_n)$ given $\{\widehat{\sigma}_i\}$ and $\mathcal{S}^{SD}_n$ where $Y_i^\star = \widehat{\sigma}_i \epsilon_i$ for $i = \overline{1,n}$.

Note that $\mathcal{S}^{SD}_n \subset \mathcal{S}^{OS}_n \subset \mathcal{S}_n$, and so $c^{SD}_\eta \leqslant c^{OS}_\eta \leqslant c^{PI}_\eta$ for any $\eta \in (0,1)$. This explains that the three methods for simulating the critical values are arranged in terms of increasing power.

4. Theory under High-Level Conditions

This section describes the high-level assumptions used in this paper and presents the main results under these assumptions.

Let C_1, C_2, ϕ, κ_1, κ_2, and κ_3 be some strictly positive constants. The size properties of the test will be obtained under the following assumptions.

A1. $\mathrm{E}[|\varepsilon_i|^{4+\phi}] \leqslant C_1$ *and* $\sigma_i \geqslant C_2$ *for all* $i = \overline{1,n}$.

This is a mild assumption on the moments of disturbances. The condition $\sigma_i \geqslant C_2$ for all $i = \overline{1,n}$ precludes the existence of super-efficient estimators.

Recall that the results in this paper are obtained for two types of estimators of σ_i. When $\widehat{\sigma}_i = \widehat{\epsilon}_i = Y_i - \widehat{f}(X_i)$ for some estimator $\widehat{f}$ of f, I will assume

A2. *(i)* $\widehat{\sigma}_i = Y_i - \widehat{f}(X_i)$ *for all* $i = \overline{1,n}$ *and (ii)* $\widehat{f}(X_i) - f(X_i) = o_p(n^{-\kappa_1})$ *uniformly over* $i = \overline{1,n}$.

This assumption is satisfied for many parametric and nonparametric estimators of f, see, in particular, subsection 3.4. When $\widehat{\sigma}_i$ is some consistent estimator of σ_i, I will assume

A3. $\widehat{\sigma}_i - \sigma_i = o_p(n^{-\kappa_2})$ *uniformly over* $i = \overline{1,n}$.

See subsection 3.4 for different available estimators. See also Section 5 and Lemma 2 in particular where Assumption A3 is proven for the local version of Rice's estimator.

A4. $(\widehat{V}(s)/V(s))^{1/2} - 1 = o_p(n^{-\kappa_3})$ *and* $(V(s)/\widehat{V}(s))^{1/2} - 1 = o_p(n^{-\kappa_2})$ *uniformly over* $s \in \mathcal{S}_n$.

This is a high-level assumption that will be verified for particular choices of the weighting functions under more primitive conditions in the next section (Lemma 3).

Let

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

I refer to A_n as a sensitivity parameter. It provides an upper bound on how much any test function depends on a particular observation. Intuitively, approximation of the distribution of the test statistic is possible only if A_n is sufficiently small.

A5. *$nA_n^3(\log p)^{7/2} = o(1)$ where $p = |\mathcal{S}_n|$, the number of elements in the set $\mathcal{S}_n$. In addition, if A2 holds, then for some $-2 < \phi_1 < \phi$, (i) $(\log p)^2/n^{(2+\phi)/(4+\phi_1)} = o(1)$, (ii) $A_n^2 n^{2/(4+\phi_1)}(\log p)^3 = o(1)$, (iii) $A_n^2(\log p)^5 = o(1)$, and (iv) $\log p/n^{\kappa_1\wedge\kappa_3} = o(1)$. Finally, if A3 is satisfied, then $\log p/n^{\kappa_2\wedge\kappa_3} = o(1)$.*

This is a key growth assumption that restricts the choice of the weighting functions and, hence, the set $\mathcal{S}_n$. Note that this condition includes p only through $\log p$, and so it allows an exponentially large (in the sample size n) number of weighting functions. Lemma 3 in the next section provides an upper bound on A_n for some choices of weighting functions, allowing me to verify this Assumption.

Let $\mathcal{M}$ be a class of models given by equation (1), regression function f, design points $\{X_i\}$, distribution of $\{\varepsilon_i\}$, weighting functions $Q(\cdot,\cdot,s)$ for $s \in \mathcal{S}_n$, and estimators $\{\widehat{\sigma}_i\}$ such that uniformly over this class, (i) Assumptions A1, A4, and A5 are satisfied, and (ii) either Assumption A2 or A3 holds.[4] For $M \in \mathcal{M}$, let $\mathrm{P}_M(\cdot)$ denote the probability under the distributions in the model M. Then

Theorem 1. *Let $P = PI$, OS, or SD. Let $\mathcal{M}_0$ denote the set of all models $M \in \mathcal{M}$ satisfying $\mathcal{H}_0$. Then*

$$\inf_{M\in\mathcal{M}_0}\mathrm{P}_M(T \leqslant c_{1-\alpha}^P) \geqslant 1-\alpha+o(1) \text{ as } n\to\infty.$$

In addition, let $\mathcal{M}_{00}$ denote the set of all models $M \in \mathcal{M}_0$ such that $f \equiv C$ for some constant C. Then

$$\sup_{M\in\mathcal{M}_{00}}\mathrm{P}(T \leqslant c_{1-\alpha}^P) = 1-\alpha+o(1) \text{ as } n\to\infty.$$

Comment 1. (i) This Theorem states that the Wild Bootstrap combined with the selection procedures developed in this paper yields valid critical values. Moreover, critical values are valid

[4]Assumptions A2, A3, and A4 contain statements of the form $Z = o_p(n^{-\kappa})$ for some random variable Z and $\kappa > 0$. I say that these assumptions hold uniformly over a class of models if for any $C > 0$, $\mathrm{P}(|Z| > Cn^{-\kappa}) = o(1)$ uniformly over this class. Note that this notion of uniformity is weaker than uniform convergence in probability. In addition, it applies to random variables defined on different probability spaces.

uniformly over the class of models $\mathcal{M}_0$. The second part of the Theorem states that the test is nonconservative in the sense that its 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 Chatterjee (2005) and Chernozhukov, Chetverikov, and Kato (2011). In particular, the validity of the bootstrap is established without refering to the asymptotic distribution of the test statistic.

(iii) Note that T has a form of U-statistic. The analysis of such statistics typically requires a preliminary Hoeffding projection. An advantage of the approximation method developed in this paper is that it applies directly to the test statistic with no need for the Hoeffding projection, which simplifies the analysis a lot.

(iv) To obtain a particular application of the general result presented in this theorem, consider a basic set of weighting functions introduced in subsection 3.2. Assume that $(\log n)^{7/2}/(nh_{\min}^3)^{1/2} \to 0$ as $n \to 0$. Then the number of weighting functions in the set $\mathcal{S}_n$ is bounded from above by some polynomial in n, and so $\log p \lesssim \log n$. Lemma 3 in the next Section then implies that Assumptions 4 and A5 hold with $\phi_1 = 0$ (under mild conditions on $K(\cdot)$ stated in Lemma 3), and so the result of Theorem 1 applies for this $\mathcal{S}_n$. Therefore, the basic set of weighting functions yields a test with the correct asymptotic size, and so it can be used for testing monotonicity. An advantage of this set is that, as will follow from Theorems 4 and 5, it gives a test with the best attainable rate of uniform consistency in the minimax sense against alternatives with regression functions that have Lipschitz-continuous first order derivatives provided that $h_{\min} \to 0$ sufficiently fast.

Let $s_l = \inf_{1 \leqslant i \leqslant \infty} X_i$ and $s_r = \sup_{1 \leqslant i \leqslant \infty} X_i$. To prove consistency of the test and to derive the rate of consistency against one-dimensional alternatives, I will also incorporate the following assumptions.

A6. *For any interval $[x, x + \Delta_x] \subset [s_l, s_r]$ there exists an integer N and a constant $C > 0$ such that for any $n \geqslant N$, $|\{i = \overline{1,n} : X_i \in [x, x + \Delta_x]\}| \geqslant Cn$.*

This Assumption often appears in the literature. Lemma 1 in the next section shows that it holds almost surely if $\{X_i\}$ is an i.i.d. sequence from some distribution satisfying mild regularity conditions.

A7. *For any interval $[x, x+\Delta_x] \subset [s_l, s_r]$ there exists an integer N and a constant $C > 0$ such that for any $n \geqslant N$, there exists $s \in \mathcal{S}_n$ satisfying (i) the support of $Q(\cdot, \cdot, s)$ is contained in $[x, x+\Delta_x]^2$, (ii) $Q(\cdot, \cdot, s)$ is bounded from above uniformly over $n = \overline{1,\infty}$, (iii) there exist nonintersecting subintervals $[x_l, x_l + \Delta_{x,l}]$ and $[x_r, x_r + \Delta_{x,r}]$ of $[x, x + \Delta_x]$ such that $Q(x_1, x_2, s) \geqslant C$ whenever $x_1 \in [x_l, x_l + \Delta_{x,l}]$ and $x_2 \in [x_r, x_r + \Delta_{x,r}]$.*

Let $\mathcal{M}_1$ be a subset of $\mathcal{M}$ consisting of all models satisfying Assumptions A6 and A7. Then

Theorem 2. *Let $P = PI$, OS, or SD. Then for any model M from the class $\mathcal{M}_1$ such that f is continuously differentiable and there exist $x_1, x_2 \in [s_l, s_r]$ such that $x_1 < x_2$ and $f(x_1) > f(x_2)$ ($\mathcal{H}_0$ is false),*

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

Comment 2. (i) This Theorem shows that the test is consistent against any fixed continuously differentiable alternative.

(ii) To compare the critical values based on the selection procedures developed in this paper with the plug-in approach (no selection procedure), assume that f is continuously differentiable and strictly increasing ($\mathcal{H}_0$ holds). Then an argument like that used in the proof of Theorem 2 shows that $\mathcal{S}^{OS}_n$ and $\mathcal{S}^{SD}_n$ will be empty w.p.a.1, which means that $\mathrm{P}\{c^{OS}_{1-\alpha} = 0\} \to 1$ and $\mathrm{P}\{c^{SD}_{1-\alpha} = 0\} \to 1$. On the other hand, $\mathrm{P}(c^{PI}_{1-\alpha} > C) \to 1$ for some $C > 0$ since each test statistic contains at least one weighting function. Moreover, under Assumption A7, it follows from the Sudakov-Chevet Theorem (see, for example, Theorem 2.3.5 in Dudley (1999)) that $\mathrm{P}(c^{PI}_{1-\alpha} > C) \to 1$ for all $C > 0$. Finally, under Assumption A9, which is stated below, it follows from the proof of lemma 2.3.15 in Dudley (1999) that $\mathrm{P}\{c^{PI}_{1-\alpha} > C\sqrt{\log n}\} \to 1$ for some $C > 0$. This explains the power improvements of one-step and stepdown approaches in comparison with the plug-in critical value.

Theorem 3. *Let $P = PI$, OS, or SD. Consider any model M from the class $\mathcal{M}_1$ such that f is continuously differentiable and there exist $x_1, x_2 \in [s_l, s_r]$ such that $x_1 < x_2$ and $f(x_1) > f(x_2)$ ($\mathcal{H}_0$ is false). Assume that for every sample size n, the true model M_n coincides with M except that the regression function has the form $f_n(\cdot) = l_n f(\cdot)$ for some sequence $\{l_n\}$ of positive numbers converging to zero. Then*

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

as long as $\log p = o(l_n^2 n)$.

Comment 3. (i) This Theorem establishes the consistency of the test against one-dimensional local alternatives, which are often used in the literature to investigate the power of the test; see, for example, Andrews and Shi (2010), Lee, Song, and Whang (2011), and the discussion in Horowitz and Spokoiny (2001).

(ii) Suppose that $\mathcal{S}_n$ consists of a basic set of weighting functions and $h_{\min} \to 0$ polynomially fast. Then $\log p \lesssim C \log n$, and so the test is consistent against one-dimensional local alternatives if $(\log n/n)^{1/2} = o(l_n)$.

(iii) Now suppose that $\mathcal{S}_n$ is a maximal subset of a basic set such that for any x_1, x_2, h satisfying $(x_1, h) \in \mathcal{S}_n$ and $(x_2, h) \in \mathcal{S}_n$, $|x_2 - x_1| > 2h$. In addition, assume that $h_{\min} \to 0$ arbitrarily

slowly. Then the test is consistent against one-dimensional local alternatives if $n^{-1/2} = o(l_n)$. In words, this test is $\sqrt{n}$-consistent against such alternatives. I note however, that the practical value of this $\sqrt{n}$-consistency is limited because there is no guarantee that for any given sample size n and given deviation from $\mathcal{H}_0$, weighting functions suitable for detecting this deviation are already included in the test statistic. In contrast, it will follow from Theorem 4 that the test based on a basic set of weighting functions does provide this guarantee.

Let $\{C_j : j = 3, ..., 8\}$ be a set of strictly positive constants such that $C_3 < C_4$, $C_5 < C_6$, and $C_7 < C_8$. Let $L > 0$, $\beta \in (0, 1]$, $k \geqslant 0$, and $h_n = (\log p/n)^{1/(2\beta+3)}$. To derive the uniform consistency rate against the classes of alternatives with Lipschitz derivatives, conditions A6 and A7 will be replaced by the following assumptions.

A 8. *There exists an integer N such that for any $n \geqslant N$ and any interval $[x_1, x_2] \subset [s_l, s_r]$ satisfying $|x_2 - x_1| \geqslant C_3 n^{-1/3}$, $C_5 n|x_2 - x_1| \leqslant |\{i = \overline{1,n} : X_i \in [x_1, x_2]\}| \leqslant C_6 n|x_2 - x_1|$.*

This Assumption is stronger than A6 but is still often imposed in the literature; see Lemma 1 for sufficient primitive conditions.

A 9. *There exists an integer N such that for any $n \geqslant N$ and any $x \in [s_l, s_r - C_4 h_n]$, there exists $s \in \mathcal{S}_n$ satisfying (i) the support of $Q(\cdot, \cdot, s)$ is contained in $[x, x + C_4 h_n]^2$, (ii) $Q(\cdot, \cdot, s)$ is bounded from above by $C_8 h_n^k$, (iii) there exist $x_l, x_r \in [x, x + C_4 h_n]$ such that $|x_r - x_l| > 2C_3 h_n$ and $Q(x_1, x_2, s) \geqslant C_7 h_n^k$ whenever $x_1 \in [x_l, x_l + C_3 h_n]$ and $x_2 \in [x_r, x_r + C_3 h_n]$.*

This Assumption is satisfied for the basic set of weighting functions if $h_{\min}$ satisfies $h_{\min} = o(\log p/n)^{1/(2\beta+3)}$. Let $f^{(1)}(\cdot)$ denote the first derivative of $f(\cdot)$.

A10. *For any $x_1, x_2 \in [s_l, s_r]$, $|f^{(1)}(x_1) - f^{(1)}(x_2)| \leqslant L|x_1 - x_2|^\beta$.*

This is a smoothness condition that requires that the regression function is sufficiently well-behaved.

Let $\mathcal{M}_2$ be the subset of $\mathcal{M}$ consisting of all models satisfying Assumptions A8, A9, and A10. Then

Theorem 4. *Let $P = PI$, OS, or SD. Consider any sequence of positive numbers $\{l_n\}$ such that $l_n \to \infty$, and let $\mathcal{M}_{2n}$ denote the subset of $\mathcal{M}_2$ consisting of all models such that the regression function f satisfies $\inf_{x \in [s_l, s_r]} f^{(1)}(x) < -l_n (\log p/n)^{\beta/(2\beta+3)}$. Then*

$$\sup_{M \in \mathcal{M}_{2n}} \mathrm{P}_M(T \leqslant c_{1-\alpha}^P) \to 0 \text{ as } n \to \infty.$$

Comment 4. (i) Theorem 4 gives the rate of uniform consistency of the test against Holder smoothness classes with parameters $(\beta + 1, L)$. Importance of *uniform* consistency against sufficiently large classes of alternatives such as Holder smoothness classes was previously emphasized

in Horowitz and Spokoiny (2001). Intuitively, it guarantees that there are no reasonable alternatives against which the test has low power if the sample size is sufficiently large.

(ii) Suppose that $\mathcal{S}_n$ consists of a basic set of weighting functions, $K(\cdot)$ is continuous and strictly positive on $(-1,+1)$, and $h_{\min}$ satisfies $h_{\min} = o(\log n/n)^{1/(2\beta+3)}$ and $(\log n)^{7/2}/(nh_{\min}^3)^{1/2} = o(1)$. Then Assumption A9 holds. In addition, it follows from Lemma 3 that Assumptions A4 and A5 are satisfied (under mild conditions on $K(\cdot)$ stated in Lemma 3), and so Theorem 4 implies that the test with this $\mathcal{S}_n$ is consistent whenever $\inf_{x\in[s_l,s_r]} f_n^{(1)}(x) < -l_n(\log n/n)^{\beta/(2\beta+3)}$ for some $l_n \to \infty$. On the other hand, it will be shown in Theorem 5 that no test can be consistent if $\inf_{x\in[s_l,s_r]} f_n^{(1)}(x) > -C(\log n/n)^{\beta/(2\beta+3)}$ for some sufficiently large $C > 0$. Therefore, the test is rate optimal in the minimax sense.

To conclude this Section, I present a Theorem that gives a lower bound on the possible rate of uniform consistency against the class $\mathcal{M}_2$ so that no test that maintains asymptotic size can have a higher rate of uniform consistency. Let $\psi = \psi(Y_1, ..., Y_n)$ be a generic test. In other words, $\psi(Y_1, ..., Y_n)$ is the probability that the test rejects upon observing the data Y_i, $i = \overline{1,n}$. Note that for any deterministic test $\psi = 0$ or 1.

Theorem 5. *For any test ψ satisfying $\mathrm{E}_M[\psi] \leqslant \alpha + o(1)$ as $n \to \infty$ for all models $M \in \mathcal{M}$ such that $\mathcal{H}_0$ holds, there exists a sequence of models $M = M_n$ belonging to the class $\mathcal{M}_2$ such that $f = f_n$ satisfies $\inf_{x\in[s_l,s_r]} f_n^{(1)}(x) < -C(\log n/n)^{\beta/(2\beta+3)}$ for some sufficiently large constant $C > 0$ and $\mathrm{E}_{M_n}[\psi] \leqslant \alpha + o(1)$ as $n \to \infty$. Here $\mathrm{E}_{M_n}[\cdot]$ denotes the expectation under the distributions of the model M_n.*

Comment 5. Combining the result of this Theorem with Comment 4-ii shows that the test based on a basic set of weighting functions and satisfying conditions of Comment 4-ii is rate optimal. In other words, no test that maintains asymptotic size can have a higher uniform consistency rate against the models with the regression function possessing the Lipschitz-continuous first order derivative.

5. Verification of High-Level Conditions

This section provides conditions that are sufficient for the Assumptions used in Section 4. First, I discuss Assumptions A6 and A8 concerning the configuration of design points $\{X_i\}$. Then I consider Assumption A3, which concerns the uniform consistency of the estimator $\widehat{\sigma}_i$ of σ_i over $i = \overline{1,n}$. Finally, I give an upper bound on the sensitivity parameter A_n and prove Assumption A4 for the case when $\mathcal{S}_n$ consists of kernel weighting functions.

Recall that the analysis in Section 4 is for nonstochastic $\{X_i\}$. Alternatively, it can be viewed as conditional on $\{X_i\}$. Suppose that $\{X_i\}$ is an i.i.d. sample from some distribution. The

Lemma below provides sufficient conditions so that Assumptions A6 and A8 hold for almost all realizations $\{X_i\}$.

Lemma 1. *Suppose that $\{X_i\}_{1\leqslant i\leqslant\infty}$ is an i.i.d. sample from the distribution P_x on $\mathbb{R}$ with the bounded support $[s_l, s_r]$. Then Assumption A6 holds for almost all realizations $\{X_i\}_{1\leqslant i\leqslant\infty}$. In addition, if P_x is absolutely continuous with respect to Lebesgue measure, and its density is bounded from above and away from zero on the support, then Assumption A8 holds for almost all realizations $\{X_i\}_{1\leqslant i\leqslant\infty}$.*[5]

Note that sufficient conditions provided by Lemma 1 for Assumption A6 allow for point masses, whereas conditions for Assumption A8 do not.

From now on, I will again assume that $\{X_i\}$ is nonstochastic. The next Lemma shows uniform consistency of the local version of Rice's estimator $\widehat{\sigma}_i$ with an explicit rate of convergence in probability.

Lemma 2. *Suppose that $\widehat{\sigma}_i$ is the local version of Rice's estimator of σ_i given in equation (7). Suppose also that (i) Assumption A1 holds, (ii) $\log n = o(n^{\phi/(4+\phi)}b_n^3)$ for some sequence $\{b_n\}$ of positive numbers converging to zero, (iii) $|J(i)| \geqslant Cnb_n$ for some $C > 0$ and all $i = \overline{1,n}$, (iv) $|f(X_i) - f(X_j)| \lesssim |X_i - X_j|$ uniformly over $i, j = \overline{1,n}$, and (v) $|\sigma_i^2 - \sigma_j^2| \lesssim |X_i - X_j|$ uniformly over $i, j = \overline{1,n}$. Then $\max_{1\leqslant i\leqslant n} |\widehat{\sigma}_i - \sigma_i| = O_p(b_n)$.*

Note that since $\phi/(4+\phi) \in (0,1)$, Assumption (iii) follows from A8, and Assumption (iv) follows from A10 as long as $\{X_i\}$ is contained in the bounded set. Lemma 2 implies that Assumption A3 holds for the local version of Rice's estimator with any κ_2 satisfying $\kappa_2 < \phi/(12+3\phi)$.

Next, I consider restrictions on the weighting functions to ensure that Assumption A4 holds and give an upper bound on the sensitivity parameter A_n.

Lemma 3. *Suppose that $\mathcal{S}_n$ consists of kernel weighting functions. In addition, suppose that (i) Assumptions A1 and A8 hold, (ii) K has the support $[-1,+1]$, is continuous, and strictly positive on the interior of its support, (iii) $x \in [s_l, s_r]$ for all $(x,h) \in \mathcal{S}_n$, (iv) $nh_{\min}^3 \to \infty$ where $h_{\min} = \min_{(x,h)\in\mathcal{S}_n} h$, and (v) $h_{\max} \leqslant (s_r - s_l)/2$ where $h_{\max} = \max_{(x,h)\in\mathcal{S}_n} h$. Then (a) $A_n \leqslant C/(nh_{\min})^{1/2}$ where C depends only on the kernel K and constants $C_1, ..., C_8$; (b) if Assumption A3 is satisfied, then Assumption A4 holds with $\kappa_3 = \kappa_2$; (c) if Assumption A2 is satisfied, then Assumption A4 holds with any $\kappa_3 < (2+\phi)/(4+\phi_1)$ for any $\phi_1 \in (-2, \phi)$ as long as $\log p = o(h_{\min} n^{1-2\kappa_3})$ and $\log p = o(h_{\min} n^{(2+\phi_1)/(4+\phi_1)-\kappa_3})$.*

[5]Recall that in section 4, s_l and s_r were defined by $s_l = \inf_{1\leqslant i\leqslant\infty} X_i$ and $s_r = \sup_{1\leqslant i\leqslant\infty} X_i$. It is easy to show that the definition given in this Lemma coincides with that definition for almost all realizations $\{X_i\}_{1\leqslant i\leqslant\infty}$.

Restrictions on the kernel K imposed in this Lemma are satisfied for most commonly used kernel functions including uniform, triangular, Epanechnikov, biweight, triweight, and tricube kernels. Note, however, that these restrictions exclude higher order kernels since those are necessarily negative at some points on their supports.

6. Models with Multivariate Covariates

Most empirical studies contain additional covariates that should be controlled for. In this section, I extend the results presented in Sections 4 and 5 to allow for this possibility. I consider cases of both partially linear and nonparametric models. For brevity, I will only consider the results concerning size properties of the test, and I will assume that σ_i is estimated by $\widehat{\sigma}_i = \widehat{\varepsilon}_i$ for all $i = \overline{1,n}$. The power properties of the test can be obtained using the arguments closely related to those used in Theorems 2, 3, and 4.

6.1. **Partially Linear Model.**

In this model, additional covariates enter the regression function as additively separable linear form. In other words, the model is given by

$$Y_i = f(X_i) + Z_i^T \beta + \varepsilon_i,\ i = 1, 2, 3, ...$$

where $\{Y_i, X_i, \varepsilon_i\}$ are defined as in the Introduction, $\{Z_i\} \subset \mathbb{R}^d$ is a sequence of nonstochastic additional covariates, and $\beta \in \mathbb{R}^d$ is a vector of coefficients. As above, the problem is to test the null hypothesis, $\mathcal{H}_0$, that $f(x)$ is nondecreasing against the alternative, $\mathcal{H}_a$, that there are x_1 and x_2 such that $x_1 < x_2$ but $f(x_1) > f(x_2)$.

An advantage of the partially linear model outlined above over the fully nonparametric model is that it does not suffer from the curse of dimensionality, which decreases the power of the test and may be a severe problem if the researcher has many additional covariates to control for. On the other hand, the partially linear model does not allow for heterogeneous effects of the factor X, which might be restrictive in some applications. It should be taken into account that the test obtained for the partially linear model will be inconsistent if this model is misspecified.

Let me now describe the test. The idea behind the test is to estimate β by $\widehat{\beta}$ and to apply the methods described in section 3 for the dataset $\{X_i, Y_i - Z_i^T\widehat{\beta}\}$. More precisely, let $\widehat{\beta}$ be a $\sqrt{n}$-consistent estimator of β. For example, one can take an estimator of Robinson (1988), which is

$$\widehat{\beta} = \left(\sum_{i=1}^n \widehat{Z}_i \widehat{Z}_i^T\right)^{-1} \left(\sum_{i=1}^n \widehat{Z}_i \widehat{Y}_i\right)$$

where $\widehat{Z}_i = Z_i - \widehat{E}[Z|X = X_i]$, $\widehat{Y}_i = Y_i - \widehat{E}[Y|X = X_i]$, and $\widehat{E}[Z|X = X_i]$ and $\widehat{E}[Y|X = X_i]$ are nonparametric estimators of $E[Z|X = X_i]$ and $E[Y|X = X_i]$ respectively; see discussion in Horowitz (2009) for a set of regularity conditions underlying $\sqrt{n}$-consistency of this estimator. Define $\tilde{Y}_i = Y_i - Z_i^T\widehat{\beta}$, and let the test statistic be $T = T(\{X_i, \tilde{Y}_i\}, \{\widehat{\sigma}_i\}, \mathcal{S}_n)$ where $\widehat{\sigma}_i = \widehat{\varepsilon}_i =$

$Y_i - \widehat{f}(X_i) - Z_i^T\widehat{\beta}$ and $\widehat{f}(X_i)$ is some estimator of $f(X_i)$, which is uniformly consistent over $i = \overline{1,n}$. The critical value for the test is simulated by one of the methods (plug-in, one-step, or stepdown) described in Section 3 using the data $\{X_i, \tilde{Y}_i\}$, estimators $\{\widehat{\sigma}_i\}$, and the set of weighting functions $\mathcal{S}_n$. As in Section 3, let $c^{PI}_{1-\alpha}$, $c^{OS}_{1-\alpha}$, and $c^{SD}_{1-\alpha}$ denote the plug-in, one-step, and stepdown critical values correspondingly.

Let $C_9 > 0$ be some constant. To obtain results for partially linear models, I will impose the following condition.

A11. *(i)* $\|Z_i\| \leqslant C_9$ *for all* $i = \overline{1,n}$*, (ii)* $\lim_{C\to\infty} \mathrm{P}(\|\widehat{\beta} - \beta\| > Cn^{-1/2}) \to 0$ *uniformly over all* n*, and (iii)* $\max_{s\in\mathcal{S}_n} \sum_{1\leqslant i,j\leqslant n} Q(X_i, X_j, s)/V(s)^{1/2} = o(\sqrt{n/\log p})$.

Let $\mathcal{M}_{PL}$ denote any set of models in $\mathcal{M}$ such that Assumptions A2 and A11 are satisfied uniformly over $\mathcal{M}_{PL}$. It follows from the proof of Lemma 3 that Assumption A11-iii is satisfied if $\mathcal{S}_n$ consists of kernel weighting functions as long as $h_{\max}$ satisfies $h_{\max} = o(1/\log p)$. The size properties of the test are given in the following theorem.

Theorem 6. *Let* $P = PI$*,* OS*, or* SD*. Let* $\mathcal{M}_{PL,0}$ *denote the set of all models* $M \in \mathcal{M}_{PL,0}$ *satisfying* $\mathcal{H}_0$*. Then*

$$\inf_{M\in\mathcal{M}_{PL,0}} \mathrm{P}_M(T \leqslant c^P_{1-\alpha}) \geqslant 1 - \alpha + o(1) \text{ as } n \to \infty.$$

In addition, let $\mathcal{M}_{PL,00}$ *denote the set of all models* $M \in \mathcal{M}_{PL,0}$ *such that* $f \equiv C$ *for some constant* C*. Then*

$$\sup_{M\in\mathcal{M}_{PL,00}} \mathrm{P}_M(T \leqslant c^P_{1-\alpha}) = 1 - \alpha + o(1) \text{ as } n \to \infty.$$

6.2. **Nonparametric Model.** In this subsection, I do not assume that the regression function is separably additive in additional covariates. Instead, I assume that the regression function has a general nonparametric form, and so the model is given by

$$Y_i = f(X_i, Z_i) + \varepsilon_i,\ i = 1, 2, 3, ...$$

where $\{X_i, Z_i\}$ is a sequence of $1 + d$ vectors of nonstochastic covariates, $\{Y_i\}$ is a sequence of scalar dependent random variables, and $\{\varepsilon_i\}$ is a sequence of unobservable scalar random variables satisfying $\mathrm{E}[\varepsilon_i] = 0$ for all $i = \overline{1,n}$.

Let S_z be some subset of $\mathbb{R}^d$. The null hypothesis, $\mathcal{H}_0$, to be tested is that for any $x_1, x_2 \in \mathbb{R}$ and $z \in S_z$, $f(x_1, z) \leqslant f(x_2, z)$ whenever $x_1 \leqslant x_2$. The alternative, $\mathcal{H}_a$, is that there are $x_1, x_2 \in \mathbb{R}$ and $z \in S_z$ such that $x_1 \leqslant x_2$ but $f(x_1, z) > f(x_2, z)$.

The choice of the set S_z is up to the researcher and has to be made depending on theoretical considerations. For example, if $S_z = \mathbb{R}^d$, then $\mathcal{H}_0$ means that the function f is increasing in the first argument for any given value of the second argument. If the researcher is interested in one

particular value, say, z_0, then she can set $S_z = z_0$, which will mean that under $\mathcal{H}_0$, the function f is increasing in the first argument when the second argument equals z_0.

The advantage of the nonparametric model studied in this subsection is that it is fully flexible and, in particular, allows for heterogeneous effects of X on Y. On the other hand, the nonparametric model suffers from the curse of dimensionality and may result in tests with low power if the researcher has many additional covariates. In this case, it might be better to consider the partially linear model studied above.

To define the test statistic, let $\mathcal{S}_n$ and $Q(\cdot, \cdot, s)$ be the same as in Section 3. Then define

$$\bar{\mathcal{S}}_n = \{(s, z) : s \in \mathcal{S}_n, z = Z_i \text{ for some } i = \overline{1, n} \text{ such that } Z_i \in S_z\},$$

and for $\bar{s} = (s, z) \in \bar{\mathcal{S}}_n$, let

$$b(\bar{s}) = (1/2) \sum_{1 \leqslant i,j \leqslant n} (Y_i - Y_j)\text{sign}(X_j - X_i)\bar{Q}(X_i, Z_i, X_j, Z_j, \bar{s})$$

be a test function where

$$\bar{Q}(X_i, Z_i, X_j, Z_j, \bar{s}) = Q(X_i, X_j, s)\bar{K}((Z_i - z)/\bar{h}(s))\bar{K}((Z_j - z)/\bar{h}(\bar{s})),$$

$\bar{K} : \mathbb{R}^d \to \mathbb{R}$ is some positive compactly supported auxiliary kernel function, and $\bar{h}(\bar{s})$, $\bar{s} \in \bar{\mathcal{S}}_n$, are auxiliary bandwidth values. Intuitively, $\bar{Q}$ is a local-in-z version of the weighting function Q. The variance of $b(\bar{s})$ is given by

$$V(\bar{s}) = \sum_{1 \leqslant i \leqslant n} \sigma_i^2 \left(\sum_{1 \leqslant j \leqslant n} \text{sign}(X_j - X_i)\bar{Q}(X_i, Z_i, X_j, Z_j, \bar{s}) \right)^2,$$

and the estimated variance is

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

Then the test statistic is

$$T = \max_{\bar{s} \in \bar{\mathcal{S}}_n} \frac{b(\bar{s})}{\sqrt{\widehat{V}(\bar{s})}}$$

Large values of T indicate that $\mathcal{H}_0$ is violated. The critical value for the test can be calculated using any of the methods described in Section 3 with the only difference being that now $\bar{Q}$, $\bar{s}$ and $\bar{\mathcal{S}}_n$ should be used instead of Q, s and $\mathcal{S}_n$, and the selection procedures choose subsets of $\bar{\mathcal{S}}_n$ instead of $\mathcal{S}_n$. Let $c_{1-\alpha}^{PI}$, $c_{1-\alpha}^{OS}$, and $c_{1-\alpha}^{SD}$ denote the plug-in, one-step, and stepdown critical values correspondingly. In addition, let

$$\bar{A}_n = \max_{\bar{s} \in \bar{\mathcal{S}}_n} \max_{1 \leqslant i \leqslant n} \left| \sum_{1 \leqslant j \leqslant n} \text{sign}(X_j - X_i)\bar{Q}(X_i, X_j, \bar{s})/(V(\bar{s}))^{1/2} \right|,$$

be a sensitivity parameter. Finally, let $\bar{p} = |\bar{\mathcal{S}}_n|$, the number of elements in the set $\bar{\mathcal{S}}_n$. Clearly, $\bar{p} \leqslant pn$ where $p = |\mathcal{S}_n|$.

Let C_{10} be some positive constant. To prove results concerning multivariate nonparametric model, I will impose the following condition.

A12. *(i)* $\mathrm{P}(|\varepsilon_i| \geqslant u) \leqslant \exp(-u^2/C_{10})$ *for all* $u \geqslant 0$ *and* $\sigma_i \geqslant C_2$ *for all* $i = \overline{1,n}$*, (ii)* $\bar{A}_n(\log \bar{p})^{7/2} = o(1)$*, (iii)* $\log \bar{p}/n^{\kappa_1 \wedge \kappa_3} = o(1)$*, (iv) for some* $-2 < \phi_1 < \infty$*,* $(\log \bar{p})^2/n^{(2+\phi)/(4+\phi_1)} = o(1)$*, and* $\bar{A}_n^2 n^{2/(4+\phi_1)}(\log \bar{p})^3 = o(1)$*, (v)* $\bar{h}(\bar{s}) \sum_{1 \leqslant i,j \leqslant n} \bar{Q}(X_i, Z_i, X_j, Z_j, \bar{s})/(V(\bar{s}))^{1/2} = o(1/\sqrt{\log \bar{p}})$ *uniformly over* $\bar{s} \in \bar{\mathcal{S}}_n$*, and (vi) the regression function* f *has uniformly bounded first order partial derivatives.*

Condition (i) of this Assumption imposes that ε_i have sub-Gaussian tails, which is stronger than Assumption A1. Conditions (ii)-(v) are of high level. To give more primitive conditions, assume that $\mathcal{S}_n$ consists of kernel weighting functions so that $\bar{s} = (s, z) = ((x, h), z)$, and that the number of points $\{X_i, Z_i\}_{1 \leqslant i \leqslant n}$ contained in each cube with the center (X_j, Z_j), $j = \overline{1,n}$, and edges $h \times \bar{h}(\bar{s})^d$ is bounded from below and from above by $\underline{C} h \bar{h}(\bar{s})^d$ and $\overline{C} h \bar{h}(\bar{s})^d$ correspondingly with some constants $0 < \underline{C} < \overline{C} < \infty$. Then $\log \bar{p} \lesssim \log n$. Let $\bar{\mathcal{S}}_{n,h} = \{(h, \bar{h}) : \bar{h} = \bar{h}((x, h), z)$ for some x and z such that $((x, h), z) \in \bar{\mathcal{S}}_n\}$. Then it follows from the proof of Lemma 3 that $\bar{A}_n \lesssim \max_{(h,\bar{h}) \in \bar{\mathcal{S}}_{n,h}} 1/(nh\bar{h}^d)^{1/2}$. Therefore, by setting ϕ_1 sufficiently large, conditions (ii)-(v) hold if $nh\bar{h}^d \to \infty$ and $nh\bar{h}^{d+2} \to 0$ polynomially fast uniformly over $(h, \bar{h}) \in \bar{\mathcal{S}}_{n,h}$.

The key difference between the multivariate case studied in this section and univariate case studied in Section 4 is that now it is not necessarily the case that $\mathrm{E}[b(\bar{s})] \leqslant 0$ under $\mathcal{H}_0$. The reason is that the values $f(x_1, z_1)$ and $f(x_2, z_2)$ are noncomparable unless $z_1 = z_2$. This yields nonvanishing bias term in the test statistic. Condition (v) of Assumption 12 ensures that this bias is asymptotically negligible relative to the concentration rate of the test statistic. The difficulty, however, is that this condition is inconsistent with $n\bar{A}_n^3(\log \bar{p})^{7/2} \to 0$ imposed in Assumption A5 (where I replaced A_n and p by their multivariate analogs $\bar{A}_n$ and $\bar{p}$). Indeed, condition $n\bar{A}_n^3(\log \bar{p})^{7/2} \to 0$ essentially requires $nh^3\bar{h}^{3d} \to \infty$, and so it contradicts to $nh\bar{h}^{d+2} \to 0$, which follows from condition (v) of A12. To deal with this problem, I impose more stringent moment condition A12-i than that used in Section 4, A1. This allows me to apply more powerful methods developed in Chernozhukov, Chetverikov, and Kato (2012) and replace $n\bar{A}_n^3(\log \bar{p})^{7/2} \to 0$ by $\bar{A}_n(\log \bar{p})^{7/2} = o(1)$; see Assumption A12-ii.

Let $\mathcal{M}_{NP}$ denote any set of models such that Assumptions A2 with $\widehat{f}(X_i)$ and $f(X_i)$ replaced by $\widehat{f}(X_i, Z_i)$ and $f(X_i, Z_i)$, A4 with s and $\mathcal{S}_n$ replaced by $\bar{s}$ and $\bar{\mathcal{S}}_n$, and A12 hold uniformly over $\mathcal{M}_{NP}$. Then

Theorem 7. *Let $P = PI$, OS, or SD. Let $\mathcal{M}_{NP,0}$ denote the set of all models $M \in \mathcal{M}_{NP}$ satisfying $\mathcal{H}_0$. Then*

$$\inf_{M \in \mathcal{M}_{NP,0}} \mathrm{P}_M(T \leqslant c^P_{1-\alpha}) \geqslant 1 - \alpha + o(1) \text{ as } n \to \infty.$$

In addition, let $\mathcal{M}_{NP,00}$ denote the set of all models $M \in \mathcal{M}_{NP,0}$ such that $f \equiv C$ for some constant C. Then

$$\sup_{M \in \mathcal{M}_{NP,00}} \mathrm{P}_M(T \leqslant c^P_{1-\alpha}) = 1 - \alpha + o(1) \text{ as } n \to \infty.$$

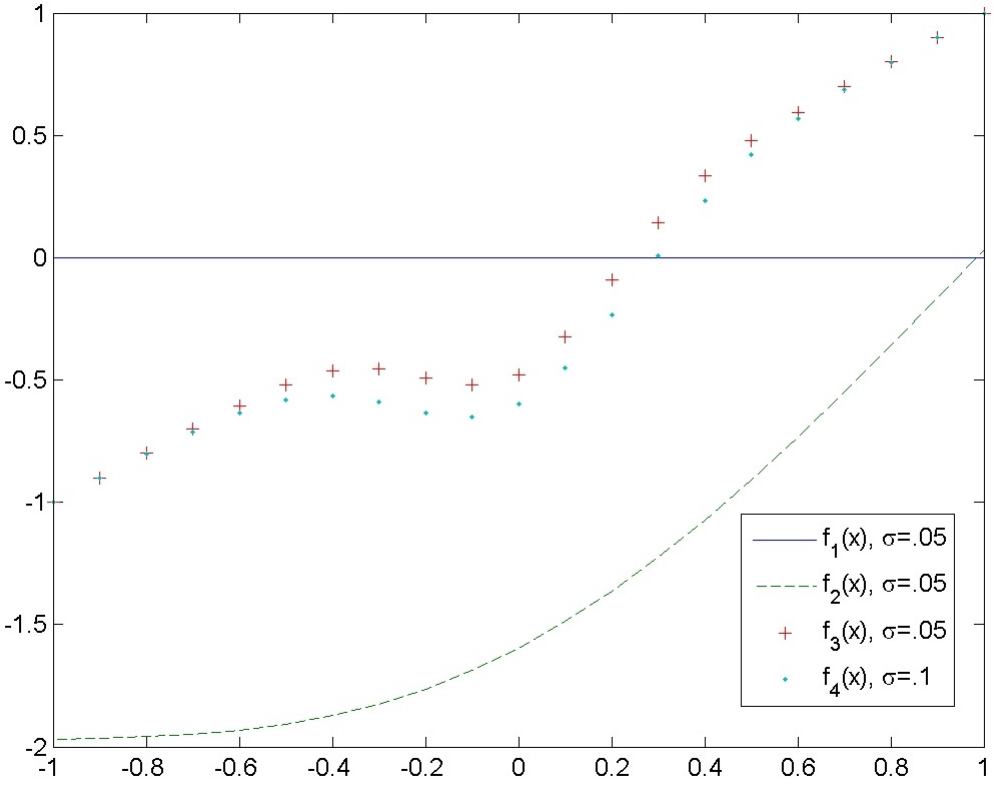
7. Monte Carlo Simulations

In this section, I provide results of a small simulation study. The aim of the simulation study is to shed some light on the size properties of the test in finite samples and to compare its power with that of other tests developed in the literature. In particular, I consider the tests of Gijbels, Hall, Jones, and Koch (2000) (GHJK), Ghosal, Sen, and van der Vaart (2000) (GSV), and Hall and Heckman (2000) (HH).

I consider samples of size $n = 100$, 200, and 500 with equidistant nonstochastic X_i's on the $[-1, 1]$ interval, and regression functions of the form $f = c_1 x - c_2 \phi(c_3 x)$ where $c_1, c_2, c_3 \geqslant 0$ and $\phi(\cdot)$ is the pdf of the standard normal distribution. I assume that $\{\varepsilon_i\}$ is a sequence of i.i.d. zero-mean random variables with standard deviation σ. Depending on the experiment, ε_i has either normal or continous uniform distribution. Four combinations of parameters are studied: (1) $c_1 = c_2 = c_3 = 0$ and $\sigma = 0.05$; (2) $c_1 = c_3 = 1$, $c_2 = 4$, and $\sigma = 0.05$; (3) $c_1 = 1$, $c_2 = 1.2$, $c_3 = 5$, and $\sigma = 0.05$; (4) $c_1 = 1$, $c_2 = 1.5$, $c_3 = 4$, and $\sigma = 0.1$. Cases 1 and 2 satisfy $\mathcal{H}_0$ whereas cases 3 and 4 do not. In case 1, the regression function is flat corresponding to the maximum of the type I error. In case 2, the regression function is strictly increasing. Cases 3 and 4 give examples of the regression functions that are mostly increasing but violate $\mathcal{H}_0$ in the small neighborhood near 0. All functions are plotted in figure 2. The parameters were chosen so that to have nontrivial rejection probability in most cases (that is, bounded from zero and from one).

Let me describe the tuning parameters for all tests that are used in the simulations. For the tests of GSV, GHJK, and HH, I tried to follow their instructions as closely as possible. For the test developed in this paper, I use kernel weighting functions with $k = 0$, $\mathcal{S}_n = \{(x, h) : x \in \{X_1, ..., X_n\}, h \in H_n\}$, and the kernel $K(x) = 0.75(1 - x^2)$ for $x \in (-1; +1)$ and 0 otherwise. I use the set of bandwidth values $H_n = \{h_{\max} u^l : h \geqslant h_{\min}, l = 0, 1, 2, ...\}$, $u = 0.5$, $h_{\max} = 1$, $h_{\min} = (0.3/n^{0.95})^{1/3}$, and the truncation parameter $\gamma = 0.01$. For the test of GSV, I use the same kernel K with the bandwidth value $h_n = n^{-1/5}$, which was suggested in their paper, and I consider their their sup-statistic. For the test of GHJK, I use their run statistic maximized over $k \in \{10(j-1) + 1 : j = 1, 2, ...0.2n\}$ (see the original paper for the explanation of the

FIGURE 2. Regression Functions Used in Simulations



notation). For the test of HH, local polynomial estimates are calculated over $r \in nH_n$ at every design point X_i. The set nH_n is chosen so that to make the results comparable with those for the test developed in this paper. Finally, I consider two versions of the test developed in this paper depending on how σ_i is estimated. More precisely, I consider the test with σ_i estimated by the Rice's method (see equation (6)), which I refer to in the table below as CS (consistent sigma), and the test with $\widehat{\sigma}_i = \widehat{\varepsilon}_i$ where $\widehat{\varepsilon}_i$ is obtained as the residual from estimating f using the series method with polynomials of order 5, 6 and 8 whenever the sample size n, is 100, 200, and 500 respectively, which I refer to in the table below as IS (inconsistent sigma).

The rejection probabilities corresponding to nominal size $\alpha = 0.1$ for all tests are presented in table 1. The results are based on 1000 simulations with 500 bootstrap repetitions in all cases excluding the test of GSV where the asymptotic critical value is used.

The results of the simulations can be summarized as follows. First, the results for normal and uniform disturbances are rather similar. The test developed in this paper with σ_i estimated using the Rice's method maintains the required size quite well (given the nonparametric structure

TABLE 1. Results of Monte Carlo Experiments

Noise	Case	Sample	Proportion of Rejections for								
			GSV	GHJK	HH	CS-PI	CS-OS	CS-SD	IS-PI	IS-OS	IS-SD
normal	1	100	.118	.078	.123	.128	.128	.128	.164	.164	.164
		200	.091	.051	.108	.114	.114	.114	.149	.149	.149
		500	.086	.078	.105	.114	.114	.114	.133	.133	.133
normal	2	100	0	.001	0	.001	.008	.008	.008	.024	.024
		200	0	.002	0	.001	.010	.010	.007	.017	.017
		500	0	.001	0	.002	.007	.007	.005	.016	.016
normal	3	100	0	.148	.033	.259	.436	.433	0	0	0
		200	.010	.284	.169	.665	.855	.861	.308	.633	.650
		500	.841	.654	.947	.982	.995	.997	.975	.995	.995
normal	4	100	.037	.084	.135	.163	.220	.223	.023	.042	.043
		200	.254	.133	.347	.373	.499	.506	.362	.499	.500
		500	.810	.290	.789	.776	.825	.826	.771	.822	.822
uniform	1	100	.109	.079	.121	.122	.122	.122	.201	.201	.201
		200	.097	.063	.109	.121	.121	.121	.160	.160	.160
		500	.077	.084	.107	.092	.092	.092	.117	.117	.117
uniform	2	100	.001	.001	0	0	.006	.007	.017	.032	.033
		200	0	0	0	.001	.010	.010	.012	.022	.024
		500	0	.003	0	.003	.011	.011	.011	.021	.021
uniform	3	100	0	.151	.038	.244	.438	.449	0	0	0
		200	.009	.233	.140	.637	.822	.839	.290	.607	.617
		500	.811	.582	.947	.978	.994	.994	.975	.990	.990
uniform	4	100	.034	.084	.137	.155	.215	.217	.024	.045	.046
		200	.197	.116	.326	.357	.473	.478	.323	.452	.456
		500	.803	.265	.789	.785	.844	.846	.782	.847	.848

Nominal Size is 0.1. GSV, GHJK, and HH stand for the tests of Ghosal, Sen, and van der Vaart (2000), Gijbels, Hall, Jones, and Koch (2000), and Hall and Heckman (2000) respectively. CS-PI, CS-OS, and CS-SD refer to the test developed in this paper with σ_i estimated using Rice's formula and plug-in, one-step, and stepdown critical values respectively. Finally, IS-PI, IS-OS, and IS-SD refer to the test developed in this paper with σ_i estimated by $\widehat{\sigma}_i = \widehat{\varepsilon}_i$ and plug-in, one-step, and stepdown critical values respectively.

of the problem) and yields size comparable with that of the GSV, GHJK, and HH tests. On the other hand, the test with $\widehat{\sigma}_i = \widehat{\varepsilon}_i$ does pretty well in terms of size only when the sample size is as large as 500. When the null hypothesis does not hold, the CS test with the stepdown

critical value yields the highest proportion of rejections in all cases. Moreover, in case 3 with the sample size $n = 200$, this test has much higher power than that of GSV, GHJK, and HH. The CS test also has higher power than that of the IS test. Finally, the table shows that the one-step critical value gives a notable improvement in terms of power in comparison with plug-in critical value. For example, in case 3 with the sample size $n = 200$, the one-step critical value gives additional 190 rejections out 1000 simulations in comparison with the plug-in critical value for the CS test and additional 325 rejections for the IS test. On the other hand, the stepdown approach gives only minor improvements over the one-step approach. Overall, the results of the simulations are consistent with the theoretical findings in this paper. In particular, selection procedures yielding one-step and stepdown critical values improve power with no size distortions. Additional simulation results are presented in the supplementary Appendix.

8. Empirical Application

In this section, I review the arguments of Ellison and Ellison (2011) on how strategic entry deterrence might yield a nonmonotone relation between market size and investment in the pharmaceutical industry and then apply the testing procedures developed in this paper to their dataset. I start with describing their theory. Then I provide the details of the dataset. Finally, I present the results.

In the pharmaceutical industry, incumbents whose patents are about to expire can use investments strategically to prevent generic entries after the expiration of the patent. In order to understand how this strategic entry deterrence influences the relation between market size and investment levels, Ellison and Ellison (2011) developed two models for an incumbent's investment. In the first model, potential entrants do not observe the incumbent's investment but they do in the second one. So, a strategic entry deterrence motive is absent in the former model but is present in the latter one. Therefore, the difference in incumbent's investment between two models is explained by the strategic entry deterrence. Ellison and Ellison showed that in the former model, the investment-market size relation is determined by a combination of direct and competition effects. The direct effect is positive if increasing the market size (holding entry probabilities fixed) raises the marginal benefit from the investment more than it raises the marginal cost of the investment. The competition effect is positive if the marginal benefit of the investment is larger when the incumbent is engaged in duopoly competition than it is when the incument is a monopolist. The equilibrium investment is increasing in market size if and only if the sum of two effects is positive. Therefore, a sufficient condition for the monotonicity of investment-market size relation is that both effects are of the same sign.[6] In the latter model, there is also a strategic entry deterrence effect. The authors noted that this effect should be

[6]An interested reader can find a more detailed discussion in the original paper.

relatively less important in small and large markets than it is in markets of intermediate size. In small markets, there are not enough profits for potential entrants, and there is no need to prevent entry. In large markets, profits are so large that no reasonable investment levels will be enough to prevent entries. As a result, strategic entry deterrence might yield a nonmonotonic relation between market size and investment no matter whether the relation in the model with no strategic entry deterrence is increasing or decreasing.

Ellison and Ellison studied three types of investment: detail advertising, journal advertising, and presentation proliferation. Detail advertising, measured as per-consumer expenditures, refers to sending representatives to doctors' offices. Since both revenues and cost of detail advertising are likely to be linear in the market size, it can be shown that the direct effect for detail advertising is zero. The competition effect is likely to be negative because detail advertising will benefit competitors as well. Therefore, it is expected that detail advertising is a decreasing function of the market size in the absence of strategic distortions. Stategic entry deterrence should decrease detail advertising for markets of intermediate size. Journal advertising is the placement of advertisements in medical journals. Journal advertising is also measured as per-consumer expenditures. The competition effect for journal advertising is expected to be negative for the same reason as for detail advertising. The direct effect, however, may be positive because the cost per potential patient is probably a decreasing function of the market size. Opposite directions of these effects make journal advertising less attractive for detecting strategic entry deterrence in comparison with detail advertising. Nevertheless, following the original paper, I assume that journal advertising is a decreasing function of the market size in the absence of strategic distortions. Presentation proliferation is selling a drug in many different forms. Since the benefits of introducing a new form is approximately proportional to the market size while the costs can be regarded as fixed, the direct effect for presentation proliferation should be positive. In addition, the competition effect is also likely to be positive because it creates a monopolistic niche for the incumbent. Therefore, presentation proliferation should be positively related to market size in the absence of strategic distortions.

The dataset consists of 63 chemical compounds, sold under 71 different brand names. All of these drugs lost their patent exclusivity between 1986 and 1992. There are four variables in the dataset: average revenue for each drug over three years before the patent expiration (this measure should be regarded as a proxy for market size), average costs of detail and journal advertising over the same time span as revenues, and a Herfindahl-style measure of the degree to which revenues are concentrated in a small number of presentations (this measure should be regarded as the inverse of presentation proliferation meaning that higher values of the measure indicate lower presentation proliferation).

Clearly, the results will depend on how I define both dependent and independent variables for the test. Following the strategy adopted in the original paper, I use log of revenues as the independent variable in all cases, and the ratio of advertising costs to revenues for detail and journal advertising and the Herfindahl-style measure for presentation proliferation as the dependent variable. The null hypothesis is that the corresponding conditional mean function is decreasing.[7]

I consider the test with kernel weighting functions with $k = 0$ or 1 and the kernel $K(x) = 0.75(1-x^2)$ for $x \in (-1, 1)$ and 0 otherwise. I use the set of bandwidth values $H_n = \{0.5; 1\}$ and the set of weighting functions $\mathcal{S}_n = \{(x, h) : x \in \{X_1, ..., X_n\}, h \in H_n\}$. Implementing the test requires estimating σ_i^2 for all $i = 1, ..., n$. Since the test based on Rice's method outperformed that with $\widehat{\sigma}_i = \widehat{\varepsilon}_i$ in the Monte Carlo simulations, I use this method in the benchmark procedure. I also check robustness of the results using the following two-step procedure. First, I obtain residuals of the OLS regression of Y on a set of transformations of X. In particular, I use polynomials in X up to the third degree (cubic polynomial). Second, squared residuals are projected onto the same polynomial in X using the OLS regression again. The resulting projections are estimators $\widehat{\sigma}_i^2$ of σ_i^2, $i = 1, ..., n$.

The results of the test are presented in table 2. The table shows the p-value of the test for each type of investment and each method of estimating σ_i^2. In the table, method 1 corresponds to estimating σ_i^2 using Rice's formula, and methods 2, 3, and 4 are based on polynomials of first, second, and third degrees respectively. Note that all methods yield similar numbers, which reassures the robustness of the results. All the methods with $k = 0$ reject the null hypothesis that journal advertising is decreasing in market size with 10% confidence level. This may be regarded as evidence that pharmaceutical companies use strategic investment in the form of journal advertising to deter generic entries. On the other hand, recall that direct and competition effects probably have different signs for journal advertising, and so rejecting the null may also be due to the fact that the direct effect dominates for some values of market size. In addition, the test with $k = 1$ does not reject the null hypothesis that journal advertising is decreasing in market size at the 10% confidence level, no matter how σ_i are estimated. No method rejects the null hypothesis in the case of detail advertising and presentation proliferation. This may be (1) because firms do not use these types of investment for strategic entry deterrence, (2) because the

[7]In the original paper, Ellison and Ellison (2011) test the null hypothesis consisting of the union of monotonically increasing and monotonically decreasing regression functions. The motivation for this modification is that increasing regression functions contradict the theory developed in the paper and, hence, should not be considered as evidence of the existence of strategic entry deterrence. On the other hand, increasing regression functions might arise if the strategic entry deterrence effect overweighs direct and competition effects even in small and large markets, which could be considered as extreme evidence of the existence of strategic entry deterrence.

TABLE 2. Incumbent Behavior versus Market Size: Monotonicity Test p-value

Method	Investment Type					
	Detail Advertising		Journal Advertising		Presentation Proliferation	
	k=0	k=1	k=0	k=1	k=0	k=1
1	.120	.111	.056	.120	.557	.661
2	.246	.242	.088	.168	.665	.753
3	.239	.191	.099	.195	.610	.689
4	.301	.238	.098	.194	.596	.695

strategic effect is too weak to yield nonmonotonicity, or (3) because the sample size is not large enough. Overall, the results are consistent with those presented in Ellison and Ellison (2011).

9. Conclusion

In this paper, 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 selection procedures. The 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. Finally, I have given tests suitable for models with multiple covariates for the first time in the literature.

The new methods have numerous applications in economics. In particular, they can be applied to test qualitative predictions of comparative statics analysis including those derived via robust comparative statics. In addition, they are useful for evaluating monotonicity assumptions, which are often imposed in economic and econometric models, and for classifying economic objects in those cases where classification includes the concept of monotonicity (for example, normal/inferior and luxury/necessity goods). Finally, these methods can be used to detect strategic behavior of economic agents that might cause nonmonotonicity in otherwise monotone relations.

The attractive properties of the new tests are demonstrated via Monte Carlo simulations. In particular, it is shown that the rejection probability of the new tests greatly exceeds that of other tests for some simulation designs. In addition, I applied the tests developed in this paper to study entry deterrence effects in the pharmaceutical industry using the dataset of Ellison and Ellison (2011). I showed that the investment in the form of journal advertising seems to be used by incumbents in order to prevent generic entries after the expiration of patents. The evidence is rather weak, though.

Appendix A. Implementation Details

In this section, I provide detailed step-by-step instructions for implementing plug-in, one-step, and stepdown critical values. The instructions are given for constructing a test of level α. In all cases, let B be a large integer denoting the number of bootstrap repetitions, and let $\{\epsilon_{i,b}\}_{i=1,b=1}^{n,B}$ be a set of independent $N(0,1)$ random variables. For one-step and stepdown critical values, let γ denote the truncation probability, which should be small relative to α.

A.1. **Plug-in Approach.**

(1) For each $b = \overline{1,B}$ and $i = \overline{1,n}$, calculate $Y_{i,b}^\star = \widehat{\sigma}_i \epsilon_{i,b}$.
(2) For each $b = \overline{1,B}$, calculate the value $T_b^\star$ of the test statistic using the sample $\{X_i, Y_{i,b}^\star\}_{i=1}^n$.
(3) Define the plug-in critical value, $c_{1-\alpha}^{PI}$, as the $(1-\alpha)$ sample quantile of $\{T_b^\star\}_{b=1}^B$.

A.2. **One-Step Approach.**

(1) For each $b = \overline{1,B}$ and $i = \overline{1,n}$, calculate $Y_{i,b}^\star = \widehat{\sigma}_i \epsilon_{i,b}$.
(2) Using the plug-in approach, simulate $c_{1-\gamma}^{PI}$.
(3) Define $\mathcal{S}_n^{OS}$ as the set of values $s \in \mathcal{S}_n$ such that $b(s)/(\widehat{V}(s))^{1/2} > -2c_{1-\gamma}^{PI}$.
(4) For each $b = \overline{1,B}$, calculate the value $T_b^\star$ of the test statistic using the sample $\{X_i, Y_{i,b}^\star\}_{i=1}^n$ and taking maximum only over $\mathcal{S}_n^{OS}$ instead of $\mathcal{S}_n$.
(5) Define the one-step critical value, $c_{1-\alpha}^{OS}$, as the $(1-\alpha)$ sample quantile of $\{T_b^\star\}_{b=1}^B$.

A.3. **Stepdown Approach.**

(1) For each $b = \overline{1,B}$ and $i = \overline{1,n}$, calculate $Y_{i,b}^\star = \widehat{\sigma}_i \epsilon_{i,b}$.
(2) Using the plug-in and one-step approaches, simulate $c_{1-\gamma}^{PI}$ and $c_{1-\gamma}^{OS}$, respectively.
(3) Denote $\mathcal{S}_n^0 = \mathcal{S}_n^{OS}$, $c^0 = c_{1-\gamma}^{OS}$, and set $l = 0$.
(4) For given value of $l \geqslant 0$, define $\mathcal{S}_n^{l+1}$ as the set of values $s \in \mathcal{S}_n^l$ such that $b(s)/(\widehat{V}(s))^{1/2} > -c_{1-\gamma}^{PI} - c^l$.
(5) For each $b = \overline{1,B}$, calculate the value $T_b^\star$ of the test statistic using the sample $\{X_i, Y_{i,b}^\star\}_{i=1}^n$ and taking the maximum only over $\mathcal{S}_n^{l+1}$ instead of $\mathcal{S}_n$.
(6) Define c^{l+1}, as the $(1-\gamma)$ sample quantile of $\{T_b^\star\}_{b=1}^B$.
(7) If $\mathcal{S}_n^{l+1} = \mathcal{S}_n^l$, then go to step (8). Otherwise, set $l = l+1$ and go to step (4).
(8) For each $b = \overline{1,B}$, calculate the value $T_b^\star$ of the test statistic using the sample $\{X_i, Y_{i,b}^\star\}_{i=1}^n$ and taking the maximum only over $\mathcal{S}_n^l$ instead of $\mathcal{S}_n$.
(9) Define $c_{1-\alpha}^{SD}$, as the $(1-\alpha)$ sample quantile of $\{T_b^\star\}_{b=1}^B$.

Appendix B. Additional Notation

I will use the following additional notation in Appendices C and D. Recall that $\{\epsilon_i\}$ is a sequence of independent $N(0,1)$ random variables that are independent of the data. Denote $e_i = \sigma_i \epsilon_i$ and $\widehat{e}_i = \widehat{\sigma}_i \epsilon_i$ for $i = \overline{1,n}$. Let

$$w_i(s) = \sum_{1 \leqslant j \leqslant n} \operatorname{sign}(X_j - X_i) Q(X_i, X_j, s),$$

$$a_i(s) = w_i(s)/(V(s))^{1/2} \text{ and } \widehat{a}_i(s) = w_i(s)/(\widehat{V}(s))^{1/2},$$

$$e(s) = \sum_{1 \leqslant i \leqslant n} a_i(s) e_i, \text{ and } \widehat{e}(s) = \sum_{1 \leqslant i \leqslant n} \widehat{a}_i(s) \widehat{e}_i,$$

$$\varepsilon(s) = \sum_{1 \leqslant i \leqslant n} a_i(s) \varepsilon_i \text{ and } \widehat{\varepsilon}(s) = \sum_{1 \leqslant i \leqslant n} \widehat{a}_i(s) \varepsilon_i,$$

$$f(s) = \sum_{1 \leqslant i \leqslant n} a_i(s) f(X_i) \text{ and } \widehat{f}(s) = \sum_{1 \leqslant i \leqslant n} \widehat{a}_i(s) f(X_i).$$

Note that $T = \max_{s \in \mathcal{S}_n} \sum_{1 \leqslant i \leqslant n} \widehat{a}_i(s) Y_i = \max_{s \in \mathcal{S}_n} (\widehat{f}(s) + \widehat{\varepsilon}(s))$. In addition, for any $\mathcal{S} \subset \mathcal{S}_n$, which may depend on the data, and all $\eta \in (0,1)$, let $c_\eta^{\mathcal{S}}$ denote the conditional η quantile of $T^\star = T(\{X_i, Y_i^\star\}, \{\widehat{\sigma}_i\}, \mathcal{S})$ given $\{\widehat{\sigma}_i\}$ and $\mathcal{S}$ where $Y_i^\star = \widehat{\sigma}_i \epsilon_i$ for $i = \overline{1,n}$, and let $c_\eta^{\mathcal{S},0}$ denote the conditional η quantile of $T^\star = T(\{X_i, Y_i^\star\}, \{\sigma_i\}, \mathcal{S})$ given $\mathcal{S}$ where $Y_i^\star = \sigma_i \epsilon_i$ for $i = \overline{1,n}$. Further, 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 $+\infty$.

Moreover, denote $\mathcal{V} = \max_{s \in \mathcal{S}_n} (V(s)/\widehat{V}(s))^{1/2}$. Let $\{\psi_n\}$ be a sequence of positive numbers converging to zero sufficiently slowly so that (i) $\log p / n^{\kappa_3} = o(\psi_n)$ (recall that by Assumption A5, $\log p / n^{\kappa_3} = o(1)$, and so such a sequence exists), (ii) uniformly over $\mathcal{S} \subset \mathcal{S}_n$ and $\eta \in (0,1)$, $\mathrm{P}(c_{\eta+\psi_n}^{\mathcal{S},0} < c_\eta^{\mathcal{S}}) = o(1)$ and $\mathrm{P}(c_{\eta+\psi_n}^{\mathcal{S}} < c_\eta^{\mathcal{S},0}) = o(1)$ (Lemma 9 establishes existence of such a sequence under Assumptions A1, A3, A4, and A5 and Lemma 13 establishes existence under Assumptions A1, A2, A4, and A5). Let

$$\mathcal{S}_n^R = \{s \in \mathcal{S}_n : f(s) > -c_{1-\gamma_n-\psi_n}^{\mathcal{S}_n,0}\}.$$

For $D = PI, OS, SD, R$, let $c_\eta^D = c_\eta^{\mathcal{S}_n^D}$ and $c_\eta^{D,0} = c_\eta^{\mathcal{S}_n^D,0}$ where $\mathcal{S}_n^{PI} = \mathcal{S}_n$. Note that $c_\eta^{PI,0}$ and $c_\eta^{R,0}$ are nonstochastic.

Finally, I denote the space of k-times continuously differentiable functions on $\mathbb{R}$ by $\mathbb{C}^k(\mathbb{R}, \mathbb{R})$. For $g \in \mathbb{C}^k(\mathbb{R}, \mathbb{R})$, the symbol $g^{(r)}$ for $r \leqslant k$ denotes the rth derivative of g, and $\|g^{(r)}\|_\infty = \sup_{t \in \mathbb{R}} |g^{(r)}(t)|$.

Appendix C. Proofs for section 4

In this Appendix, I first prove a sequence of auxiliary lemmas (subsection C.1). Then I present the proofs of the theorems stated in section 4 (subsection C.2).

C.1. **Auxiliary Lemmas.**

Lemma 4. $\mathrm{E}[\max_{s\in\mathcal{S}_n}|e(s)|]\lesssim(\log p)^{1/2}$.

Proof. Note that by construction, $e(s)$ is distributed as a $N(0,1)$ random variable, and $|\mathcal{S}_n|=p$. So, the result follows from lemma 2.2.2 in Van der Vaart and Wellner (1996). □

Lemma 5. *Uniformly over $\mathcal{S}\subset\mathcal{S}_n$ and $\Delta>0$, $\sup_{t\in\mathbb{R}}\mathrm{P}(\max_{s\in\mathcal{S}}e(s)\in(t,t+\Delta))\lesssim\Delta(\log p)^{1/2}$. In particular, for any $(\eta,\delta)\in(0,1)^2$ and $\mathcal{S}\subset\mathcal{S}_n$, $c^{\mathcal{S},0}_{\eta+\delta}-c^{\mathcal{S},0}_{\eta}\geqslant C\delta/(\log p)^{1/2}$ for some constant $C>0$.*

Proof. The first claim follows by combining Lemma 4 in this paper and Theorem 1 in Chernozhukov, Chetverikov, and Kato (2011). The second claim follows from the result in the first claim. □

Lemma 6. *There exists a constant $C>0$ such that for all $\mathcal{S}\subset\mathcal{S}_n$, $\eta\in(0,1)$, and $t\in\mathbb{R}$,*

$$c^{\mathcal{S},0}_{\eta-C|t|\log p/(1-\eta)}\leqslant c^{\mathcal{S},0}_{\eta}(1+t)\leqslant c^{\mathcal{S},0}_{\eta+C|t|\log p/(1-\eta)}.$$

Proof. Recall that $c^{\mathcal{S},0}_{\eta}$ is the η quantile of $\max_{s\in\mathcal{S}}e(s)$, and so combining Lemma 4 and Markov inequality shows that $c^{\mathcal{S},0}_{\eta}\lesssim(\log p)^{1/2}/(1-\eta)$. Therefore, Lemma 5 gives

$$c^{\mathcal{S},0}_{\eta+C|t|\log p/(1-\eta)}-c^{\mathcal{S},0}_{\eta}\geqslant C|t|(\log p)^{1/2}/(1-\eta)\geqslant|t|c^{\mathcal{S},0}_{\eta}.$$

The lower bound follows similarly. □

Lemma 7. *Under Assumption A1, uniformly over $\mathcal{S}\subset\mathcal{S}_n$, $\beta>0$, and $g\in\mathbb{C}^3(\mathbb{R},\mathbb{R})$,*

$$|\mathrm{E}[g(\max_{s\in\mathcal{S}}\varepsilon(s))-g(\max_{s\in\mathcal{S}}e(s))]|\lesssim\|g^{(1)}\|_\infty\log p/\beta+nA_n^3(\|g^{(3)}\|_\infty+\beta\|g^{(2)}\|_\infty+\beta^2\|g^{(1)}\|_\infty).$$

Proof. This lemma is closely related to theorem 1.5 in Chatterjee (2005) but improves the bound. For $x=(x_1,...,x_n)\in\mathbb{R}^n$, let $x(s)=\sum_{1\leqslant i\leqslant n}a_i(s)x_i$. Let $F_\beta:\mathbb{R}^n\to\mathbb{R}$ be given by

$$F_\beta(x)=\beta^{-1}\log\left(\sum_{s\in\mathcal{S}}\exp(\beta x(s))\right)$$

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

$$\max_{s\in\mathcal{S}}x(s)=\beta^{-1}\log\left(\exp(\beta\max_{s\in\mathcal{S}}x(s))\right)\leqslant\beta^{-1}\log\left(\sum_{s\in\mathcal{S}}\exp(\beta x(s))\right)$$
$$\leqslant\beta^{-1}\log\left(p\exp(\beta\max_{s\in\mathcal{S}}x(s))\right)\leqslant\beta^{-1}\log p+\max_{s\in\mathcal{S}}x(s),$$

and so

$$|\max_{s\in\mathcal{S}}x(s)-F_\beta(x)|\leqslant\beta^{-1}\log p.$$

Therefore,

$$|\mathrm{E}[g(\max_{s\in\mathcal{S}}\varepsilon(s)) - g(\max_{s\in\mathcal{S}}e(s))]| \leqslant 2\|g^{(1)}\|_\infty \log p/\beta + |\mathrm{E}[g(F_\beta(\varepsilon) - g(F_\beta(e)]|$$

where $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ and $e = (e_1, ..., e_n)$. For $j = \overline{1,n}$, let $x^j = (\varepsilon_1, ..., \varepsilon_j, e_{j+1}, ..., e_n)$, and let $x^0 = e$. Then

$$|\mathrm{E}[g(F_\beta(\varepsilon) - g(F_\beta(e)]| \leqslant \sum_{1\leqslant j\leqslant n} |\mathrm{E}[g(F_\beta(x^j)) - g(F_\beta(x^{j-1}))]|.$$

Let $m : \mathbb{R}^n \to \mathbb{R}$ be given by $m(x) = g(F_\beta(x))$ for all $x \in \mathbb{R}^n$, and let $\partial_j^k m$ denote the k-th partial derivative of m with respect to the argument j. Then a Taylor expansion yields

$$\begin{aligned} g(F_\beta(x^j)) - g(F_\beta(x^{j-1})) &= \partial_j m(x^{j,0})(\varepsilon_j - e_j) + \partial_j^2 m(x^{j,0})(\varepsilon_j^2 - e_j^2)/2 \\ &\quad + (\partial_j^3 m(x^{j,0,\varepsilon})\varepsilon_j^3 - \partial_j^3 m(x^{j,0,e})e_j^3)/6 \end{aligned}$$

for some n-vectors $x^{j,0,\varepsilon}$ and $x^{j,0,e}$ where $x^{j,0} = (\varepsilon_1, ..., \varepsilon_{j-1}, 0, e_{j+1}, ..., e_n)$. Since ε_j and e_j are jointly independent of $x^{j,0}$ and $\mathrm{E}[\varepsilon_j] = \mathrm{E}[e_j] = 0$, $\mathrm{E}[\partial_j m(x^{j,0})(\varepsilon_j - e_j)] = 0$. In addition, $\mathrm{E}[\partial_j^2 m(x^{j,0})(\varepsilon_j^2 - e_j^2)/2] = 0$ because $\mathrm{E}[\varepsilon_j^2] = \mathrm{E}[e_j^2] = \sigma_j^2$. So, by assumption A1,

$$|\mathrm{E}[g(F_\beta(x^j)) - g(F_\beta(x^{j-1}))]| \lesssim \sup_{x\in\mathbb{R}^n} |\partial_j^3 m(x)|.$$

Finally, simple algebra shows that

$$\sup_{x\in\mathbb{R}^n} |\partial_j^3 m(x)| \lesssim A_n^3(\|g^{(3)}\|_\infty + \beta\|g^{(2)}\|_\infty + \beta^2\|g^{(1)}\|_\infty).$$

Combining presented inequalities gives the asserted claim. □

Lemma 8. *Under Assumptions A1 and A5, uniformly over $\mathcal{S} \subset \mathcal{S}_n$ and $\eta \in (0,1)$,*

$$\mathrm{P}(\max_{s\in\mathcal{S}}\varepsilon(s) \leqslant c_\eta^{\mathcal{S},0}) = \eta + o(1) \text{ and } \mathrm{P}(\max_{s\in\mathcal{S}}(-\varepsilon(s)) \leqslant c_\eta^{\mathcal{S},0}) = \eta + o(1).$$

Proof. By Assumption A5, $nA_n^3(\log p)^{7/2} \to 0$. Therefore, I can choose a sequence $\{\xi_n\}$ of positive numbers such that $\xi_n \to \infty$ and $\xi_n^5 nA_n^3(\log p)^{7/2} \to 0$. Let $g : \mathbb{R} \to [0,1]$ be a function from the class $\mathbb{C}^3(\mathbb{R},\mathbb{R})$ satisfying $g(x) = 1$ for $x \leqslant 0$ and $g(x) = 0$ for $x \geqslant 1$. Let $g_n(x) = g(\xi_n(\log p)^{1/2}(x - c_\eta^{\mathcal{S},0}))$. Finally, let $\beta_n = \xi_n^2(\log p)^{3/2}$. Then

$$\|g_n^{(1)}\|_\infty \log p/\beta_n \lesssim \xi_n(\log p)^{3/2}/\beta_n \to 0.$$

In addition,

$$nA_n^3(\|g^{(3)}\|_\infty + \beta_n\|g^{(2)}\|_\infty + \beta_n^2\|g^{(1)}\|_\infty) \lesssim \xi_n^5 nA_n^3(\log p)^{7/2} \to 0.$$

Therefore, applying Lemma 7 yields

$$\mathrm{E}[g_n(\max_{s\in\mathcal{S}}\varepsilon(s)) - g_n(\max_{s\in\mathcal{S}}e(s))] \to 0.$$

Finally, Lemma 5 gives

$$\begin{aligned}&\mathrm{P}(\max_{s\in\mathcal{S}}\varepsilon(s)\leqslant c_\eta^{\mathcal{S},0})\leqslant \mathrm{E}[g_n(\max_{s\in\mathcal{S}}\varepsilon(s))]\leqslant \mathrm{E}[g_n(\max_{s\in\mathcal{S}}e(s))]+o(1)\\&\leqslant \mathrm{P}(\max_{s\in\mathcal{S}}e(s)\leqslant c_\eta^{\mathcal{S},0}+1/(\xi_n(\log p)^{1/2}))+o(1)\leqslant \eta+o(1).\end{aligned}$$

The upper bound follows similarly. Combining the lower and the upper bounds gives the first result. The second result follows similarly. Note that all convergence statements hold uniformly over $\mathcal{S}\subset\mathcal{S}_n$ and $\eta\in(0,1)$. □

Lemma 9. *Under Assumptions A1, A3, A4, and A5, there exists a sequence $\{\psi_n\}$ of positive numbers converging to zero such that uniformly over $\mathcal{S}\subset\mathcal{S}_n$ and $\eta\in(0,1)$, $\mathrm{P}(c_{\eta+\psi_n}^{\mathcal{S},0}<c_\eta^{\mathcal{S}})=o(1)$ and $\mathrm{P}(c_{\eta+\psi_n}^{\mathcal{S}}<c_\eta^{\mathcal{S},0})=o(1)$.*

Proof. Denote

$$T^{\mathcal{S}}=\max_{s\in\mathcal{S}}\widehat{e}(s)=\max_{s\in\mathcal{S}}\sum_{1\leqslant i\leqslant n}\widehat{a}_i(s)\widehat{\sigma}_i\epsilon_i \text{ and } T^{\mathcal{S},0}=\max_{s\in\mathcal{S}}e(s)=\max_{s\in\mathcal{S}}\sum_{1\leqslant i\leqslant n}a_i(s)\sigma_i\epsilon_i.$$

Note that $c_\eta^{\mathcal{S}}$ is the conditional η quantile of $T^{\mathcal{S}}$ given $\{\widehat{\sigma}_i\}$ and $c_\eta^{\mathcal{S},0}$ is the unconditional η quantile of $T^{\mathcal{S},0}$. In addition, denote

$$p_1=\max_{s\in\mathcal{S}}|e(s)|\max_{s\in\mathcal{S}}|1-(V(s)/\widehat{V}(s))^{1/2}|,$$

$$p_2=\max_{s\in\mathcal{S}}|\sum_{1\leqslant i\leqslant n}a_i(s)(\widehat{\sigma}_i-\sigma_i)\epsilon_i|\max_{s\in\mathcal{S}}(V(s)/\widehat{V}(s))^{1/2}.$$

Then $|T^{\mathcal{S}}-T^{\mathcal{S},0}|\leqslant p_1+p_2$. Combining Lemma 4 and Assumption A4 gives

$$p_1=o_p((\log p)^{1/2}n^{-\kappa_3}).$$

Consider p_2. Conditional on $\{\widehat{\sigma}_i\}$, $(\widehat{\sigma}_i-\sigma_i)\epsilon_i$ is distributed as a $N(0,(\widehat{\sigma}_i-\sigma_i)^2)$ random variable, and so applying the argument like that in Lemma 4 conditional on $\{\widehat{\sigma}_i\}$ and using Assumptions A1 and A3 gives

$$\max_{s\in\mathcal{S}}|\sum_{1\leqslant i\leqslant n}a_i(s)(\widehat{\sigma}_i-\sigma_i)\epsilon_i|=o_p((\log p)^{1/2}n^{-\kappa_2}).$$

Since $\max_{s\in\mathcal{S}}(V(s)/\widehat{V}(s))^{1/2}\to_p 1$ by assumption A4, this implies that

$$p_2=o_p((\log p)^{1/2}n^{-\kappa_2}).$$

Therefore, $T^{\mathcal{S}}-T^{\mathcal{S},0}=o_p((\log p)^{1/2}n^{-\kappa_2\wedge\kappa_3})$, and so there exists a sequence $\{\tilde{\psi}_n\}$ of positive numbers converging to zero such that

$$\mathrm{P}(|T^{\mathcal{S}}-T^{\mathcal{S},0}|>(\log p)^{1/2}n^{-\kappa_2\wedge\kappa_3})=o(\tilde{\psi}_n).$$

Hence,

$$\mathrm{P}(\mathrm{P}(|T^{\mathcal{S}}-T^{\mathcal{S},0}|>(\log p)^{1/2}n^{-\kappa_2\wedge\kappa_3}|\{\widehat{\sigma}_i\})>\tilde{\psi}_n)\to 0.$$

Let A_n denote the event that

$$\mathrm{P}(|T^{\mathcal{S}} - T^{\mathcal{S},0}| > (\log p)^{1/2} n^{-\kappa_2 \wedge \kappa_3} | \{\widehat{\sigma}_i\}) \leqslant \tilde{\psi}_n.$$

I will take $\psi_n = \tilde{\psi}_n + C(\log p) n^{-\kappa_2 \wedge \kappa_3}$ for a constant C that is larger than that in the statement of Lemma 5. By assumption A5, $\psi_n \to 0$. Then note that

$$\mathrm{P}(T^{\mathcal{S},0} \leqslant c_\eta^{\mathcal{S},0} | \{\widehat{\sigma}_i\}) \geqslant \eta \text{ and } \mathrm{P}(T^{\mathcal{S}} \leqslant c_\eta^{\mathcal{S}} | \{\widehat{\sigma}_i\}) \geqslant \eta$$

for any $\eta \in (0,1)$. So, on A_n,

$$\begin{aligned} \eta + \tilde{\psi}_n &\leqslant \mathrm{P}(T^{\mathcal{S},0} \leqslant c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n} | \{\widehat{\sigma}_i\}) \\ &\leqslant \mathrm{P}(T^{\mathcal{S}} \leqslant c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n} + (\log p)^{1/2} n^{-\kappa_2 \wedge \kappa_3} | \{\widehat{\sigma}_i\}) + \tilde{\psi}_n \\ &\leqslant \mathrm{P}(T^{\mathcal{S}} \leqslant c^{\mathcal{S},0}_{\eta+\psi_n} | \{\widehat{\sigma}_i\}) + \tilde{\psi}_n \end{aligned}$$

where the last line uses Lemma 5. Therefore, on A_n, $c_\eta^{\mathcal{S}} \leqslant c^{\mathcal{S},0}_{\eta+\psi_n}$, i.e. $\mathrm{P}(c^{\mathcal{S},0}_{\eta+\psi_n} < c_\eta^{\mathcal{S}}) = o(1)$. The second claim follows similarly. □

Lemma 10. *Let $c_\eta^{\mathcal{S},1}$ denote the conditional η quantile of $T^{\mathcal{S},1} = \max_{s \in \mathcal{S}} \sum_{1 \leqslant i \leqslant n} a_i(s) \varepsilon_i \epsilon_i$ given $\{\varepsilon_i\}$. Let Assumptions A1, A2, and A5 hold. Then there exists a sequence $\{\tilde{\psi}_n\}$ of positive numbers converging to zero such that $\mathrm{P}(c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n} < c_\eta^{\mathcal{S},1}) = o(1)$ and $\mathrm{P}(c^{\mathcal{S},1}_{\eta+\tilde{\psi}_n} < c_\eta^{\mathcal{S},0}) = o(1)$ uniformly over $\mathcal{S} \subset \mathcal{S}_n$ and $\eta \in (0,1)$.*

Proof. I will invoke the following result recently obtained by Chernozhukov, Chetverikov, and Kato (2012).

Lemma 11. *Let Z^1 and Z^2 be zero-mean Gaussian p-vectors with covariances Σ^1 and Σ^2 correspondingly. Then for any $g \in \mathbb{C}^2(\mathbb{R}, \mathbb{R})$,*

$$|\mathrm{E}[g(\max_{1 \leqslant j \leqslant p} Z_j^1) - g(\max_{1 \leqslant j \leqslant p} Z_j^2)]| \leqslant \|g^{(2)}\|_\infty \Delta_\Sigma / 2 + 2 \|g^{(1)}\|_\infty \sqrt{2 \Delta_\Sigma \log p}$$

where $\Delta_\Sigma = \max_{1 \leqslant j,k \leqslant p} |\Sigma^1_{jk} - \Sigma^2_{jk}|$.

Proof. See Theorem 1 in Chernozhukov, Chetverikov, and Kato (2012). □

Let $Z^1 = \{\sum_{1 \leqslant i \leqslant n} a_i(s) \varepsilon_i \epsilon_i\}_{s \in \mathcal{S}}$ and $Z^2 = \{\sum_{1 \leqslant i \leqslant n} a_i(s) \sigma_i \epsilon_i\}_{s \in \mathcal{S}}$. Conditional on $\{\varepsilon_i\}$, these are zero-mean p-vectors with covariances Σ^1 and Σ^2 given by

$$\Sigma^1_{s_1 s_2} = \sum_{1 \leqslant i \leqslant n} a_i(s_1) a_i(s^2) \varepsilon_i^2 \text{ and } \Sigma^2_{s_1 s_2} = \sum_{1 \leqslant i \leqslant n} a_i(s_1) a_i(s^2) \sigma_i^2$$

Let $\Delta_\Sigma = \max_{s_1, s_2 \in \mathcal{S}} |\Sigma^1_{s_1 s_2} - \Sigma^2_{s_1 s_2}|$. The following Lemma will be helpful.

Lemma 12. $(\log p)^2 \Delta_\Sigma = o_p(1)$.

Proof. Let $u = u_n = n^{1/(4+\phi_1)}$ where ϕ_1 is given in Assumption A5. Let $\tilde{\varepsilon}_i = \varepsilon_i 1\{|\varepsilon_i| \leqslant u\}$, and let $\tilde{\sigma}_i^2 = \mathrm{E}[\tilde{\varepsilon}_i^2]$. It follows from assumption A1 that $\mathrm{P}(\max_{1 \leqslant i \leqslant n} |\tilde{\varepsilon}_i - \varepsilon_i| = 0) \to 1$. In addition,

$$0 \leqslant \sigma_i^2 - \tilde{\sigma}_i^2 = \mathrm{E}[\varepsilon_i^2 1\{|\varepsilon_i| > u\}] \leqslant \mathrm{E}[|\varepsilon_i|^{4+\phi} 1\{|\varepsilon_i| > u\}/u^{2+\phi}] \lesssim 1/u^{2+\phi}$$

uniformly over $i = \overline{1,n}$, and so

$$\begin{aligned}
&(\log p)^2 \left| \sum_{1 \leqslant i \leqslant n} a_i(s_1) a_i(s_2) (\sigma_i^2 - \tilde{\sigma}_i^2) \right| \lesssim (\log p)^2 \sum_{1 \leqslant i \leqslant n} |a_i(s_1) a_i(s_2)|/u^{2+\phi} \\
&\lesssim_{(1)} (\log p)^2 \sum_{1 \leqslant i \leqslant n} |a_i(s_1) a_i(s_2) \sigma_i^2|/u^{2+\phi} \leqslant_{(2)} (\log p)^2 \sqrt{\sum_{1 \leqslant i \leqslant n} a_i(s_1)^2 \sigma_i^2 \sum_{1 \leqslant i \leqslant n} a_i(s_2)^2 \sigma_i^2}/u^{2+\phi} \\
&=_{(3)} (\log p)^2/u^{2+\phi} =_{(4)} o(1)
\end{aligned}$$

where (1) is by Assumption A1, (2) is by Holder inequality, (3) is because $\sum_{1 \leqslant i \leqslant n} a_i(s)^2 \sigma_i^2 = 1$ by construction, and (4) is by Assumption A5. Therefore,

$$(\log p)^2 \Delta_\Sigma = (\log p)^2 \max_{s_1, s_2 \in \mathcal{S}} \left| \sum_{1 \leqslant i \leqslant n} a_i(s_1) a_i(s_2) (\tilde{\varepsilon}_i^2 - \tilde{\sigma}_i^2) \right| + o_p(1).$$

Note that $|a_i(s_1) a_i(s_2) (\tilde{\varepsilon}_i^2 - \tilde{\sigma}_i^2)| \leqslant 2A_n^2 u^2$. In addition,

$$\mathrm{E}\left[\sum_{1 \leqslant i \leqslant n} a_i(s_1)^2 a_i(s_2)^2 (\tilde{\varepsilon}_i^2 - \tilde{\sigma}_i^2)^2 \right] \lesssim A_n^2$$

uniformly over $s_1, s_2 \in \mathcal{S}$ since $\mathrm{E}[(\tilde{\varepsilon}_i^2 - \tilde{\sigma}_i^2)^2] \leqslant \mathrm{E}[\tilde{\varepsilon}_i^4] \leqslant \mathrm{E}[\varepsilon_i^4] \lesssim 1$ by Assumption A1. Hence, applying Bernstein inequality (see, for example, Lemma 2.2.9 in Van der Vaart and Wellner (1996)) gives for some $C > 0$,

$$\mathrm{P}\left((\log p)^2 \left| \sum_{1 \leqslant i \leqslant n} a_i(s_1) a_i(s_2) (\tilde{\varepsilon}_i^2 - \tilde{\sigma}_i^2) \right| > t \right) \leqslant 2 \exp\left(- \frac{t^2}{C(\log p)^4 A_n^2 + C(\log p)^2 t A_n^2 u^2} \right)$$

for any $t > 0$, and so by the union bound,

$$\begin{aligned}
&\mathrm{P}\left(\max_{s_1, s_2 \in \mathcal{S}} (\log p)^2 \left| \sum_{1 \leqslant i \leqslant n} a_i(s_1) a_i(s_2) (\tilde{\varepsilon}_i^2 - \tilde{\sigma}_i^2) \right| > t \right) \\
&\leqslant 2 \exp\left(2 \log p - \frac{t^2}{C(\log p)^4 A_n^2 + C(\log p)^2 t A_n^2 u^2} \right).
\end{aligned}$$

The result follows because Assumption A5 implies that $\log p = o(1/((\log p)^4 A_n^2))$ and $\log p = o(1/((\log p)^2 A_n^2 u^2))$. □

It follows from Lemma 12 that there exists a sequence $\{\tilde{\psi}_n\}$ of positive numbers converging to zero such that

$$(\log p)^2 \Delta_\Sigma = o_p(\tilde{\psi}_n^4). \tag{11}$$

Let $g \in \mathbb{C}^2(\mathbb{R}, \mathbb{R})$ be a function satisfying $g(t) = 1$ for $t \leqslant 0$, $g(t) = 0$ for $t \geqslant 1$, and $g(t) \in [0,1]$ for $t \in [0,1]$, and let $g_n(t) = g((t - c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n/2})/(c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n} - c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n/2}))$. Then

$$\|g_n^{(1)}\|_\infty \lesssim 1/(c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n} - c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n/2}) \lesssim (\log p)^{1/2}/\tilde{\psi}_n,$$
$$\|g_n^{(2)}\|_\infty \lesssim 1/(c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n} - c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n/2})^2 \lesssim (\log p)/\tilde{\psi}_n^2.$$

Applying Lemma 11 gives

$$D_n = |\mathrm{E}[g_n(\max_{s\in\mathcal{S}} Z_s^1) - g_n(\max_{s\in\mathcal{S}} Z_s^2)|\{\varepsilon_n\}]| \lesssim (\log p)\Delta_\Sigma/\tilde{\psi}_n^2 + (\log p)(\Delta_\Sigma)^{1/2}/\tilde{\psi}_n = o_p(\tilde{\psi}_n) \quad (12)$$

by equation (11). Note that $\max_{s\in\mathcal{S}} Z_j^1 = T^{\mathcal{S},1}$ and, using the notation of the proof of Lemma 9, $\max_{s\in\mathcal{S}} Z_s^2 = T^{\mathcal{S},0}$. Then

$$\begin{aligned}&\mathrm{P}(T^{\mathcal{S},1} \leqslant c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n}|\{\varepsilon_i\}) \geqslant_{(1)} \mathrm{E}[g_n(T^{\mathcal{S},1})|\{\varepsilon_i\}] \geqslant_{(2)} \mathrm{E}[g_n(T^{\mathcal{S},0})|\{\varepsilon_i\}] - D_n \\ &\geqslant_{(3)} \mathrm{P}(T^{\mathcal{S},0} \leqslant c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n}|\{\varepsilon_i\}) - D_n =_{(4)} \mathrm{P}(T^{\mathcal{S},0} \leqslant c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n/2}) - D_n \geqslant \eta + \tilde{\psi}_n/2 - D_n \quad (13)\end{aligned}$$

where (1) and (3) are by construction of the function g_n, (2) is by equation (12), and (4) is because $T^{\mathcal{S},0}$ and $c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n/2}$ are jointly independent of $\{\varepsilon_i\}$. Finally, note that the right hand side of line (13) is bounded from below by η w.p.a.1. This implies that $\mathrm{P}(c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n} < c^{\mathcal{S},1}_\eta) = o(1)$, which is the first asserted claim. The second claim of the Lemma follows similarly. □

Lemma 13. *Under Assumptions A1, A2, A4, and A5, there exists exists a sequence $\{\psi_n\}$ of positive numbers converging to zero such that uniformly over $\mathcal{S} \subset \mathcal{S}_n$ and $\eta \in (0,1)$, $\mathrm{P}(c^{\mathcal{S},0}_{\eta+\psi_n} < c^{\mathcal{S}}_\eta) = o(1)$ and $\mathrm{P}(c^{\mathcal{S}}_{\eta+\psi_n} < c^{\mathcal{S},0}_\eta) = o(1)$.*[8]

Proof. Lemma 10 established that

$$\mathrm{P}(c^{\mathcal{S},0}_{\eta+\tilde{\psi}_n} < c^{\mathcal{S},1}_\eta) = o(1) \text{ and } \mathrm{P}(c^{\mathcal{S},1}_{\eta+\tilde{\psi}_n} < c^{\mathcal{S},0}_\eta) = o(1).$$

Therefore, it suffices to show that

$$\mathrm{P}(c^{\mathcal{S}}_{\eta+\widehat{\psi}_n} < c^{\mathcal{S},1}_\eta) = o(1) \text{ and } \mathrm{P}(c^{\mathcal{S},1}_{\eta+\widehat{\psi}_n} < c^{\mathcal{S}}_\eta) = o(1).$$

for some sequence $\{\widehat{\psi}_n\}$ of positive numbers converging to zero. Denote

$$p_1 = \max_{s\in\mathcal{S}} |\sum_{1\leqslant i\leqslant n} a_i(s)\varepsilon_i\epsilon_i| \max_{s\in\mathcal{S}} |1 - (V(s)/\widehat{V}(s))^{1/2}|,$$

$$p_2 = \max_{s\in\mathcal{S}} |\sum_{1\leqslant i\leqslant n} a_i(s)(\widehat{\sigma}_i - \varepsilon_i)\epsilon_i| \max_{s\in\mathcal{S}} (V(s)/\widehat{V}(s))^{1/2}.$$

Note that $|T^{\mathcal{S}} - T^{\mathcal{S},1}| \leqslant p_1 + p_2$ and that by Lemmas 4 and 10, $\max_{s\in\mathcal{S}} |\sum_{1\leqslant i\leqslant n} a_i(s)\varepsilon_i\epsilon_i| = O_p((\log p)^{1/2})$. Therefore, the result follows by the argument similar to that used in the proof of Lemma 9 since $\widehat{\sigma}_i - \varepsilon_i = o_p(n^{-\kappa_1})$ by assumption A2. □

[8]Note that Lemmas 9 and 13 provide the same results under two different methods for estimating σ_i.

Lemma 14. *Let Assumptions A1, A4, and A5 hold. In addition, let either Assumption A2 or A3 hold. 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. Suppose that $\mathcal{S}_n^R \backslash \mathcal{S}_n^{SD} \neq \emptyset$. Then there exists the smallest integer l such that $\mathcal{S}_n^R \backslash \mathcal{S}_n^l \neq \emptyset$, and so $\mathcal{S}_n^R \subset \mathcal{S}_n^{l-1}$ (if $l = 1$, let $\mathcal{S}_n^0 = \mathcal{S}_n$). Therefore, $c_{1-\gamma_n}^R \leqslant c_{1-\gamma_n}^{l-1}$. It follows that there exists an element s of $\mathcal{S}_n^R$ such that

$$\widehat{f}(s) + \widehat{\varepsilon}(s) \leqslant -c_{1-\gamma_n}^{PI} - c_{1-\gamma_n}^{l-1} \leqslant -c_{1-\gamma_n}^{PI} - c_{1-\gamma_n}^{R},$$

and so

$$\begin{aligned}
\mathrm{P}(\mathcal{S}_n^R \backslash \mathcal{S}_n^{SD} \neq \emptyset) \quad &\leqslant \quad \mathrm{P}(\min_{s \in \mathcal{S}_n^R}(\widehat{f}(s) + \widehat{\varepsilon}(s)) \leqslant -c_{1-\gamma_n}^{PI} - c_{1-\gamma_n}^{R}) \\
&\leqslant_{(1)} \quad \mathrm{P}((\min_{s \in \mathcal{S}_n^R}(f(s) + \varepsilon(s))\mathcal{V} \leqslant -c_{1-\gamma_n}^{PI} - c_{1-\gamma_n}^{R}) \\
&\leqslant_{(2)} \quad \mathrm{P}((\min_{s \in \mathcal{S}_n^R}(f(s) + \varepsilon(s))\mathcal{V} \leqslant -c_{1-\gamma_n-\psi_n}^{PI,0} - c_{1-\gamma_n-\psi_n}^{R,0}) + o(1) \\
&\leqslant_{(3)} \quad \mathrm{P}((\min_{s \in \mathcal{S}_n^R}(\varepsilon(s) - c_{1-\gamma_n-\psi_n}^{PI,0})\mathcal{V} \leqslant -c_{1-\gamma_n-\psi_n}^{PI,0} - c_{1-\gamma_n-\psi_n}^{R,0}) + o(1) \\
&=_{(4)} \quad \mathrm{P}((\max_{s \in \mathcal{S}_n^R}(-\varepsilon(s)) \geqslant c_{1-\gamma_n-\psi_n}^{PI,0}(1/\mathcal{V} - 1) + c_{1-\gamma_n-\psi_n}^{R,0}/\mathcal{V}) + o(1) \\
&\leqslant_{(5)} \quad \mathrm{P}((\max_{s \in \mathcal{S}_n^R}(-\varepsilon(s)) \geqslant c_{1-\gamma_n-\psi_n}^{R,0}/\mathcal{V} - C(\log p)^{1/2} n^{-\kappa_3}/(\gamma_n + \psi_n)) + o(1) \\
&\leqslant_{(6)} \quad \mathrm{P}((\max_{s \in \mathcal{S}_n^R}(-\varepsilon(s)) \geqslant c_{1-\gamma_n-\psi_n-C(\log p)n^{-\kappa_3}/(\gamma_n+\psi_n)}^{R,0}) + o(1) \\
&\leqslant_{(7)} \quad \gamma_n + \psi_n + C(\log p)n^{-\kappa_3}/(\gamma_n + \psi_n) + o(1) =_{(8)} \gamma_n + o(1)
\end{aligned}$$

where (1) follows from the definitions of $\widehat{f}(s)$ and $\widehat{\varepsilon}(s)$, (2) is by the definition of ψ_n, (3) is by the definition of $\mathcal{S}_n^R$, (4) is rearrangement, (5) is by Lemma 4 and Assumption A4, (6) is by Lemma 5, (7) is by Lemma 8, and (8) follows from the definition of ψ_n again. The first asserted claim follows. The second claim follows from the fact that $\mathcal{S}_n^{SD} \subset \mathcal{S}_n^{OS}$. □

Lemma 15. *Let Assumptions A1, A4, and A5 hold. In addition, let either Assumption A2 or A3 hold. Then* $\mathrm{P}(\max_{s \in \mathcal{S}_n \backslash \mathcal{S}_n^R}(\widehat{f}(s) + \widehat{\varepsilon}(s)) \leqslant 0) \geqslant 1 - \gamma_n + o(1)$.

Proof. The result follows from

$$\begin{aligned}
&\mathrm{P}(\max_{s \in \mathcal{S}_n \backslash \mathcal{S}_n^R}(\widehat{f}(s) + \widehat{\varepsilon}(s)) \leqslant 0) = \mathrm{P}(\max_{s \in \mathcal{S}_n \backslash \mathcal{S}_n^R}(f(s) + \varepsilon(s)) \leqslant 0) \geqslant_{(1)} \mathrm{P}(\max_{s \in \mathcal{S}_n \backslash \mathcal{S}_n^R} \varepsilon(s) \leqslant c_{1-\gamma_n-\psi_n}^{PI,0}) \\
&\geqslant \mathrm{P}(\max_{s \in \mathcal{S}_n} \varepsilon(s) \leqslant c_{1-\gamma_n-\psi_n}^{PI,0}) =_{(2)} 1 - \gamma_n - \psi_n + o(1) =_{(3)} 1 - \gamma_n + o(1)
\end{aligned}$$

where (1) follows from the definition of $\mathcal{S}_n^R$, (2) is by Lemma 8, and (3) is by the definition of ψ_n. □

C.2. **Proofs of Theorems.**

Proof of Theorem 1. Note that

$$\begin{aligned}
&\mathrm{P}(T \leqslant c^P_{1-\alpha}) = \mathrm{P}(\max_{s\in\mathcal{S}_n}(\widehat{f}(s)+\widehat{\varepsilon}(s)) \leqslant c^P_{1-\alpha}) \geqslant_{(1)} \mathrm{P}(\max_{s\in\mathcal{S}^R_n}(\widehat{f}(s)+\widehat{\varepsilon}(s)) \leqslant c^P_{1-\alpha}) - \gamma_n + o(1)\\
&\geqslant_{(2)} \mathrm{P}(\max_{s\in\mathcal{S}^R_n}(\widehat{f}(s)+\widehat{\varepsilon}(s)) \leqslant c^R_{1-\alpha}) - 2\gamma_n + o(1) \geqslant_{(3)} \mathrm{P}(\max_{s\in\mathcal{S}^R_n}\widehat{\varepsilon}(s) \leqslant c^R_{1-\alpha}) - 2\gamma_n + o(1)\\
&\geqslant_{(4)} \mathrm{P}(\max_{s\in\mathcal{S}^R_n}\varepsilon(s)\mathcal{V} \leqslant c^{R,0}_{1-\alpha-\psi_n}) - 2\gamma_n + o(1) =_{(5)} \mathrm{P}(\max_{s\in\mathcal{S}^R_n}\varepsilon(s) \leqslant c^{R,0}_{1-\alpha-\psi_n}/\mathcal{V}) - 2\gamma_n + o(1)\\
&\geqslant_{(6)} \mathrm{P}(\max_{s\in\mathcal{S}^R_n}\varepsilon(s) \leqslant c^{R,0}_{1-\alpha-\psi_n}(1-n^{-\kappa_3})) - 2\gamma_n + o(1)\\
&\geqslant_{(7)} \mathrm{P}(\max_{s\in\mathcal{S}^R_n}\varepsilon(s) \leqslant c^{R,0}_{1-\alpha-\psi_n-C(\log p)n^{-\kappa_3}/(\alpha+\psi_n)}) - 2\gamma_n + o(1)\\
&=_{(8)} 1-\alpha-\psi_n - C(\log p)n^{-\kappa_3}/(\alpha+\psi_n) - 2\gamma_n + o(1) =_{(9)} 1-\alpha+o(1)
\end{aligned}$$

where (1) follows from Lemma 15, (2) is by Lemma 14, (3) is because under $\mathcal{H}_0$ $\widehat{f}(s) \leqslant 0$, (4) follows from the definitions of $\widehat{\varepsilon}(s)$ and ψ_n, (5) is rearrangement, (6) is by Assumption A4, (7) is by Lemma 6, (8) is by Lemma 8, and (9) is by the definitions of ψ_n and γ_n. The first asserted claim follows.

In addition, when f is identically constant,

$$\begin{aligned}
&\mathrm{P}(T \leqslant c^P_{1-\alpha}) =_{(1)} \mathrm{P}(\max_{s\in\mathcal{S}_n}\widehat{\varepsilon}(s) \leqslant c^P_{1-\alpha}) \leqslant_{(2)} \mathrm{P}(\max_{s\in\mathcal{S}_n}\widehat{\varepsilon}(s) \leqslant c^{PI}_{1-\alpha}) + \gamma_n + o(1)\\
&\leqslant_{(3)} \mathrm{P}(\max_{s\in\mathcal{S}_n}\widehat{\varepsilon}(s) \leqslant c^{PI,0}_{1-\alpha+\psi_n}) + \gamma_n + o(1) \leqslant_{(4)} \mathrm{P}(\max_{s\in\mathcal{S}_n}\varepsilon(s) \leqslant c^{PI,0}_{1-\alpha+\psi_n}(1+n^{-\kappa_3})) + \gamma_n + o(1)\\
&\leqslant_{(5)} \mathrm{P}(\max_{s\in\mathcal{S}_n}\varepsilon(s) \leqslant c^{PI,0}_{1-\alpha+\psi_n+C(\log p)n^{-\kappa_3}/(\alpha-\psi_n)}) + \gamma_n + o(1) \leqslant_{(6)} 1-\alpha+o(1)
\end{aligned}$$

where (1) follows from the fact that $\widehat{f}(s) = 0$ whenever f is identically constant, (2) follows from Lemma 14, (3) is by the definition of ψ_n, (4) is by Assumption A4, (5) is by Lemma 6, and (6) is from Lemma 8 and the definitions of γ_n and ψ_n. The second asserted claim follows. □

Proof of Theorem 2. Suppose that $f(x_2) < f(x_1)$ for some $x_1, x_2 \in [s_l, s_r]$ satisfying $x_2 > x_1$. By the mean value theorem, there exists $x_0 \in (x_1, x_2)$ satisfying

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

Therefore, $f'(x_0) < 0$. Since $f'(\cdot)$ is continuous, $f'(x) < f'(x_0)/2$ for any $x \in [x_0 - \Delta_x, x_0 + \Delta_x]$ for some $\Delta_x > 0$. Take $s = s_n \in \mathcal{S}_n$ as in Assumption A7 applied to the interval $[x_0 - \Delta_x, x_0 + \Delta_x]$. By Assumptions A1 and A7-(ii), $V(s) \leqslant Cn^3$. In addition, combining Assumptions A6, A7-(i) and A7-(iii) gives

$$\sum_{1\leqslant i,j\leqslant n}(f(X_i) - f(X_j))\mathrm{sign}(X_j - X_i)Q(X_i, X_j, s) \geqslant Cn^2 \tag{14}$$

for some $C > 0$. Further, since $\sum_{1 \leqslant i \leqslant n} a_i(s)^2 \sigma_i^2 = 1$, Assumption A1 implies $A_n \geqslant C/n^{1/2}$ for some $C > 0$, and so Assumption A5 gives $\log p = o(n)$. Therefore,

$$\begin{aligned}
\mathrm{P}(T \leqslant c^P_{1-\alpha}) &\leqslant_{(1)} \mathrm{P}(T \leqslant c^{PI}_{1-\alpha}) \leqslant_{(2)} \mathrm{P}(T \leqslant c^{PI,0}_{1-\alpha+\psi_n}) + o(1) \\
&\leqslant_{(3)} \mathrm{P}(T \leqslant C(\log p)^{1/2}) + o(1) \leqslant_{(4)} \mathrm{P}(\widehat{f}(s) + \widehat{\varepsilon}(s) \leqslant C(\log p)^{1/2}) + o(1) \\
&\leqslant_{(5)} \mathrm{P}(f(s) + \varepsilon(s) \leqslant C(\log p)^{1/2}(1 + n^{-\kappa_3})) + o(1) \\
&\leqslant_{(6)} \mathrm{P}(f(s) + \varepsilon(s) \leqslant 2C(\log p)^{1/2}) + o(1) \\
&\leqslant_{(7)} \mathrm{P}(\varepsilon(s) \leqslant 2C(\log p)^{1/2} - Cn^{1/2}) + o(1) \leqslant_{(8)} \mathrm{P}(\varepsilon(s) \leqslant -Cn^{1/2}) + o(1) \\
&\leqslant_{(9)} \mathrm{P}(\max_{s \in \mathcal{S}_n}(-\varepsilon(s)) \geqslant Cn^{1/2}) + o(1) \\
&\leqslant_{(10)} \mathrm{P}(\max_{s \in \mathcal{S}_n}(-\varepsilon(s)) \geqslant c^{PI,0}_{1-C(\log p/n)^{1/2}}) + o(1) \leqslant_{(11)} C(\log p/n)^{1/2} + o(1) = o(1)
\end{aligned}$$

where (1) follows from $\mathcal{S}^P_n \subset \mathcal{S}^{PI}_n$, (2) is by the definition of ψ_n, (3) is by Lemma 4, (4) is since $T = \max_{s \in \mathcal{S}_n}(\widehat{f}(s) + \widehat{\varepsilon}(s))$, (5) is by Assumption A4, (6) is obvious, (7) is by equation (14) and that $V(s) \leqslant Cn^3$, (8) follows from $\log p = o(n)$, (9) is obvious, (10) is by Lemma 4 and Markov inequality, and (11) follows by Lemma 8. The result follows. □

Proof of Theorem 3. The proof follows from an argument similar to that used in the proof of Theorem 2 with equation (14) replaced by

$$\sum_{1 \leqslant i,j \leqslant n} (f(X_i) - f(X_j))\mathrm{sign}(X_j - X_i)Q(X_i, X_j, s) \geqslant Cl_n n^2$$

and condition $\log p = o(n)$ replaced by $\log p = o(l_n^2 n)$. □

Proof of Theorem 4. Since $\inf_{x \in [s_l, s_r]} f^{(1)}(x) < -l_n(\log p/n)^{\beta/(2\beta+3)}$, for sufficiently large n, there exists an interval $[x_{n,1}, x_{n,2}] \subset [s_l, s_r]$ such that $|x_{n,2} - x_{n,1}| = C_4 h_n$ and for all $x \in [x_{n,1}, x_{n,2}]$, $f^{(1)}(x) < -l_n(\log p/n)^{\beta/(2\beta+3)}/2$. Take $s = s_n \in \mathcal{S}_n$ as in Assumption A9 applied to the interval $[x_{n,1}, x_{n,2}]$ By Assumptions A1, A8, and A9-(ii), $V(s) \leqslant C(nh)^3 h_n^{2k}$. In addition, combining Assumptions A8, A9-(i), and A9-(iii),

$$\sum_{1 \leqslant i,j \leqslant n} (f(X_i) - f(X_j))\mathrm{sign}(X_j - X_i)Q(X_i, X_j, s) \geqslant l_n C h_n^{1+\beta+k}(nh)^2$$

for some $C > 0$, and so $f(s) \geqslant l_n h_n^{1+\beta}(nh)^{1/2}$. From this point, since $\log p = o(l_n^2 h_n^{2\beta+3} n)$, the argument like that used in the proof of Theorem 2 yields the result. □

Proof of Theorem 5. Consider any sequence $\{X_i\}$ satisfying Assumption A8. Let $h = h_n = C_0(\log n/n)^{1/(2\beta+3)}$ for sufficiently small $C_0 > 0$. Let $L = [(s_r - s_l)/(4h)]$ where $[x]$ is the largest integer smaller or equal than x. For $l = \overline{1, L}$, let $x_l = 4h(l-1)$ and define $f_l : [s_l, s_r] \to \mathbb{R}$ by $f_l(s_l) = 0$, $f_l^{(1)}(x) = 0$ if $x \leqslant x_l$, $f_l^{(1)}(x) = -L(x - x_l)^\beta$ if $x \in (x_l, x_l + h]$, $f_l^{(1)}(x) = -L(x_l + 2h - x)^\beta$ if $x \in (x_l + h, x_l + 2h]$, $f_l^{(1)}(x) = L(x - x_l - 2h)^\beta$ if $x \in (x_l + 2h, x_l + 3h]$, $f_l^{(1)}(x) = L(x_l + 4h - x)^\beta$

if $x \in (x_l + 3h, x_l + 4h]$ and $f_l^{(1)}(x) = 0$ otherwise. In addition, let $f_0(x) = 0$ for all $x \in [s_l, s_r]$. Finally, let $\{\varepsilon\}$ be a sequence of independent $N(0,1)$ random variables.

For $l = \overline{0, L}$, consider a model $M_l = M_{n,l}$ with the sequence of design points $\{X_i\}$, the regression function f_l, and the noise $\{\varepsilon_i\}$. Note that M_0 belongs to $\mathcal{M}$ and satisfies $\mathcal{H}_0$. In addition, for $l \geqslant 1$, M_l belongs to $\mathcal{M}_2$, does not satisfy $\mathcal{H}_0$, and, moreover, has $\inf_{x \in [s_l, s_r]} f_l^{(1)}(x) < -C(\log n/n)^{\beta/(2\beta+3)}$.

Consider any test $\psi = \psi(Y_1, ..., Y_n)$ such that $\mathrm{E}_{M_0}[\psi] \leqslant \alpha + o(1)$. Then following the argument from Dumbgen and Spokoiny (2001) gives

$$\inf_{M \in \mathcal{M}_2} \mathrm{E}_M[\psi] - \alpha \leqslant \min_{1 \leqslant l \leqslant L} \mathrm{E}_{M_l}[\psi] - \mathrm{E}_{M_0}[\psi] + o(1) \leqslant \sum_{1 \leqslant l \leqslant L} \mathrm{E}_{M_l}[\psi]/L - \mathrm{E}_{M_0}[\psi] + o(1)$$
$$= \sum_{1 \leqslant l \leqslant L} \mathrm{E}_{M_0}[\psi \rho_l]/L - \mathrm{E}_{M_0}[\psi] + o(1) = \sum_{1 \leqslant l \leqslant L} \mathrm{E}_{M_0}[\psi(\rho_l - 1)]/L + o(1)$$
$$\leqslant \mathrm{E}_{M_0}[\psi |\sum_{1 \leqslant l \leqslant L} \rho_l/L - 1|] + o(1) \leqslant \mathrm{E}_{M_0}[|\sum_{1 \leqslant l \leqslant L} \rho_l/L - 1|] + o(1)$$

where ρ_l is the likelihood ratio of observing $\{Y_i\}_{1 \leqslant i \leqslant n}$ under the models M_l and M_0. Further,

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

where $\omega_{n,l} = (\sum_{1 \leqslant i \leqslant n} f_l(X_i)^2)^{1/2}$ and $\xi_{n,l} = \sum_{1 \leqslant i \leqslant n} Y_i f_l(X_i)/\omega_{n,l}$. Note that under the model M_0, $\{\xi_{n,l}\}_{1 \leqslant l \leqslant L}$ is a sequence of independent $N(0,1)$ random variables. In addition, by the construction of the functions f_l and since Assumption 8 holds, $\omega_{n,l} \leqslant Cn^{1/2}h^{\beta+3/2} = C(\log n)^{1/2}$ where C can be made arbitrarily small by selecting sufficiently small C_0. Therefore,

$$\mathrm{E}_{M_0}[|\sum_{1 \leqslant l \leqslant L} \rho_l/L - 1|] \leqslant (\mathrm{E}_{M_0}[(\sum_{1 \leqslant l \leqslant L} \rho_l/L - 1)^2])^{1/2} \leqslant (\sum_{1 \leqslant l \leqslant L} \mathrm{E}_{M_0}[\rho_l^2/L^2])^{1/2}$$
$$\leqslant (\sum_{1 \leqslant l \leqslant L} \mathrm{E}_{M_0}[\exp(2\omega_{n,l}\xi_{n,l} - \omega_{n,l}^2)/L^2])^{1/2} \leqslant (\sum_{1 \leqslant l \leqslant L} \exp(\omega_{n,l}^2)/L^2)^{1/2}$$
$$\leqslant (\exp(C^2 \log n - \log L))^{1/2} = \exp((C^2 \log n - \log L)/2) = o(1)$$

because C is arbitrarily small and $\log n \lesssim \log L$. Therefore, $\inf_{M \in \mathcal{M}_2} \mathrm{E}_M[\psi] \leqslant \alpha + o(1)$, and so the result follows. □

Appendix D. Proofs for Section 5

Proof of Lemma 1. Let X be a random variable distributed according to the law P_x. Then $\{X_i\}$ is an i.i.d. sample from the distribution of X. Let $I_i = 1\{X_i \in [x_1, x_2]\}$ for $[x_1, x_2] \subset [s_l, s_r]$. Then $\mathrm{E}[I_i] = p = \mathrm{P}_x([x_1, x_2]) > 0$. By Hoeffding inequality (see, for example, Appendix B in

Pollard (1984)),

$$\mathrm{P}(\sum_{1\leqslant i\leqslant n} I_i < pn/2) = \mathrm{P}(\sum_{1\leqslant i\leqslant n} (I_i - \mathrm{E}[I_i]) < -pn/2) \leqslant \exp(-p^2n^2/(8n)) = \exp(-p^2n/8).$$

Since $\sum_{1\leqslant n\leqslant\infty} \exp(-p^2n/8) < \infty$, the first asserted claim follows by the Borel-Cantelli Lemma.

To prove the second claim, let $U_n = [1/(C_3 n^{-1/3})]+1$ where $[\cdot]$ denotes the largest integer that is smaller or equal than the quantity inside the brackets. Let $s_l = x_{n,0} < x_{n,1} < ... < x_{n,U_n} = s_r$ where $x_{n,u} - x_{n,u-1} = (s_r - s_l)/U_n = h_{n0}$. It clearly suffices to show that for almost all realizations $\{X_i\}$ there exists an integer N such that for any $n \geqslant N$,

$$C_5 n h_{n0} \leqslant |\{i = \overline{1,n} : X_i \in [x_{n,u-1}, x_{n,u}]\}| \leqslant C_6 n h_{n0}$$

for all $u = \overline{1, U_n}$. Let $p_{n,u} = P_x([x_{n,u-1}, x_{n,u}])$. Then by Assumption, there exist constants $\underline{C}$ and $\overline{C}$ such that $\underline{C} h_{n0} \leqslant p_{n,u} \leqslant \overline{C} h_{n0}$. Let $I_{i,n,u} = 1\{X_i \in [x_{n,u-1}, x_{n,u}]\}$. Then $\mathrm{E}[I_{i,n,u}] = \mathrm{E}[I_{i,n,u}^2] = p_{n,u}$, and so Bernstein inequality gives

$$\begin{aligned}
&\mathrm{P}(\sum_{1\leqslant i\leqslant n} I_{i,n,u} > 2\overline{C} n h_{n0}) \leqslant \mathrm{P}(\sum_{1\leqslant i\leqslant n} (I_{i,n,u} - \mathrm{E}[I_{i,n,u}]) > \overline{C} n h_{n0}) \\
&\leqslant \exp(-\overline{C}^2 n^2 h_{n0}^2/(2\overline{C} n h_{n0} + 4\overline{C} n h_{n0}/3)) \leqslant \exp(-C n h_{n0})
\end{aligned}$$

for some $C > 0$. Then by the union bound,

$$\begin{aligned}
&\mathrm{P}(\max_{1\leqslant u\leqslant U_n} \sum_{1\leqslant i\leqslant n} I_{i,n,u} - 2\overline{C} n h_{n0} \geqslant 0) \leqslant \sum_{1\leqslant u\leqslant U_n} \mathrm{P}(\sum_{1\leqslant i\leqslant n} I_{i,n,u} \geqslant 2\overline{C} n h_{n0}) \\
&\leqslant \exp(C(\log(1/h_{n0}) - n h_{n0})) \leqslant \exp(-C n^{1/2}).
\end{aligned}$$

Since $\sum_{1\leqslant n\leqslant\infty} \exp(-Cn^{1/2}) < \infty$, Borel-Cantelli Lemma implies that for almost all realizations $\{X_i\}$ there exists N such that for any $n \geqslant N$, $|\{i = \overline{1,n} : X_i \in [x_{n,u-1}, x_{n,u}]\}| \leqslant C_6 n h_{n0}$ for all $u = \overline{1, U_n}$ as long as $C_6 > 2\overline{C}$. The lower bound follows similarly. Combining these bounds gives the second asserted claim. □

Proof of Lemma 2. For $B > 0$, let $u_{n,B} = B n^{1/(4+\phi)}$. In addition, define $A_{n,B}$ as the event that $\{\max_{1\leqslant i\leqslant n} |\varepsilon_i| \leqslant u_{n,B}\}$. Note that $\mathrm{P}(A_{n,B}) \to 1$ as $B \to \infty$ uniformly over $n = \overline{1,\infty}$ by Assumption A1. Further,

$$\begin{aligned}
&\mathrm{E}[|\widehat{\sigma}_i^2 - \sigma_i^2||A_{n,B}] \leqslant_{(1)} \mathrm{E}[|\widehat{\sigma}_i^2 - \sigma_i^2|]/\mathrm{P}(A_{n,B}) \leqslant_{(2)} (\mathrm{E}[(\widehat{\sigma}_i^2 - \sigma_i^2)^2])^{1/2}/\mathrm{P}(A_{n,B}) \\
&\leqslant_{(3)} \left(\mathrm{E}[(\sum_{j\in J(i): j+1\in J(i)} (Y_{j+1} - Y_j)^2/(2|J(i)|) - \sigma_i^2)^2]\right)^{1/2} /\mathrm{P}(A_{n,B}) \\
&\lesssim_{(4)} (1/|J(i)|^{1/2} + b_n)/\mathrm{P}(A_{n,B}) \lesssim_{(5)} (1/(nb_n)^{1/2} + b_n)/\mathrm{P}(A_{n,B}) \lesssim_{(6)} b_n/\mathrm{P}(A_{n,B})
\end{aligned}$$

where (1) follows from the definition of conditional expectation, (2) is by Jensen inequality, (3) is by the definition of the local version of Rice's estimator, (4) is by Assumptions (iv) and (v), (5) follows from Assumption (iii), and (6) is from Assumption (ii). In addition, exponential

concentration inequality for functions with bounded differences (see, for example, Theorem 12 in Boucheron, Bousquet, and Lugosi (2004)) gives for any $t > 0$,

$$\mathrm{P}(||\widehat{\sigma}_i^2 - \sigma_i^2| - \mathrm{E}[|\widehat{\sigma}_i^2 - \sigma_i^2||A_{n,B}]| > t|A_{n,B}) \leqslant 2\exp(-C|J(i)|t^2/u_{n,B}^4)$$

for some $C > 0$, and so using the fact that $|J(i)| > Cnb_n$, the union bound with $t = b_n$ yields

$$\mathrm{P}(\max_{1\leqslant i\leqslant n} |\widehat{\sigma}_i^2 - \sigma_i^2| > Cb_n(1 + 1/\mathrm{P}(A_{n,B}))|A_{n,B}) \lesssim \exp(\log n - n^{\phi/(4+\phi)}b_n^3) = o(1)$$

for any given $B > 0$ where the last equality follows by Assumption (ii). Therefore, $\max_{1\leqslant i\leqslant n} |\widehat{\sigma}_i^2 - \sigma_i^2| = O_p(b_n)$. Finally, since σ_i is bounded from above and away from zero uniformly over i by Assumption A1, it follows that $\max_{1\leqslant i\leqslant n} |\widehat{\sigma}_i - \sigma_i| = O_p(b_n)$, which is the asserted claim. □

Proof of Lemma 3. Let $s = (x, h) \in \mathcal{S}_n$. Since $h \leqslant (s_r - s_l)/2$, I have either $s_l + h \leqslant x$ or $x + h \leqslant s_r$. I will consider the former case. The result for the latter case follows from the same argument. Let $\bar{C}_1 \in (0, 1)$. Since the kernel K is continuous and strictly positive on its support, $\min_{t\in[0,\bar{C}_1]} K(t) > 0$. In addition, since K is bounded, I can find a constant $\bar{C}_2 \in (0, 1)$ such that

$$2C_6(1 - \bar{C}_2)^{k+1} \max_{t\in[-1,-\bar{C}_2]} K(t) \leqslant C_5\bar{C}_2^k\bar{C}_1 \min_{t\in[0,\bar{C}_1]} K(t) \tag{15}$$

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

Then for $X_i \in [x - (1 + \bar{C}_2)h/2, x - \bar{C}_2h]$,

$$\begin{aligned}
&\sum_{1\leqslant j\leqslant n} \mathrm{sign}(X_j - X_i)|X_j - X_i|^k K((X_j - x)/h) \\
&\geqslant_{(1)} \sum_{1\leqslant j\leqslant n: X_j\geqslant x} (\bar{C}_2h)^k K((X_j - x)/h) - \sum_{1\leqslant j\leqslant n: X_j\leqslant x-\bar{C}_2h} ((1 - \bar{C}_2)h)^k K((X_j - x)/h) \\
&\geqslant_{(2)} (\bar{C}_2h)^k C_5\bar{C}_1nh \min_{t\in[0,\bar{C}_1]} K(t) - ((1 - \bar{C}_2)h)^k C_6(1 - \bar{C}_2)nh \max_{t\in[-1,-\bar{C}_2]} K(t) \\
&\geqslant_{(3)} (\bar{C}_2h)^k C_5\bar{C}_1nh \min_{t\in[0,\bar{C}_1]} K(t)/2 \geqslant_{(4)} Cnh^{k+1}
\end{aligned}$$

for some $C > 0$ that depends only on $\{C_j : j = \overline{3, 8}\}$, $\bar{C}_1$, $\bar{C}_2$, and the kernel K where (1) follows from the fact that $X_i \leqslant x - \bar{C}_2h$, (2) is by Assumption A8, (3) is by equation (15), and (4) is because $\min_{t\in[0,\bar{C}_1]} K(t) > 0$. Then for $M_n(x, h) = \{i = \overline{1, n} : X_i \in [x - (1 + \bar{C}_2)h/2, x - \bar{C}_2h]\}$,

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

and so $V(s) \geqslant C(nh)^3 h^{2k}$ by Assumptions A1 and A8 where $C > 0$ does not depend on (x, h). Therefore, claim (a) follows since

$$\left| \sum_{1 \leqslant j \leqslant n} \text{sign}(X_j - X_i) Q(X_i, X_j, s) \right| \leqslant Cnh^{k+1}.$$

Further, under Assumption A3,

$$\begin{aligned}
&|\widehat{V}(s) - V(s)| \\
&\leqslant \sum_{1 \leqslant i \leqslant n} |\widehat{\sigma}_i^2 - \sigma_i^2| K((X_i - x)/h)^2 \left(\sum_{1 \leqslant j \leqslant n} \text{sign}(X_j - X_i) |X_j - X_i|^k K((X_j - x)/h) \right)^2 \\
&\leqslant \max_{1 \leqslant i \leqslant n} |\widehat{\sigma}_i^2 - \sigma_i^2| \sum_{1 \leqslant i \leqslant n} K((X_i - x)/h)^2 \left(\sum_{1 \leqslant j \leqslant n} \text{sign}(X_j - X_i) |X_j - X_i|^k K((X_j - x)/h) \right)^2,
\end{aligned}$$

and so $|\widehat{V}(s) - V(s)| \leqslant C(nh)^3 h^{2k} o_p(n^{-\kappa_2})$. Combining this bound with the lower bound for $V(s)$ established above shows that under Assumption A3, $|\widehat{V}(s)/V(s) - 1| = o_p(n^{-\kappa_2})$, and so

$$\begin{aligned}
|(\widehat{V}(s)/V(s))^{1/2} - 1| &= o_p(n^{-\kappa_2}), \\
|(V(s)/\widehat{V}(s))^{1/2} - 1| &= o_p(n^{-\kappa_2})
\end{aligned}$$

uniformly over $\mathcal{S}_n$, which is the asserted claim (b).

To prove the last claim, note that

$$|\widehat{V}(s) - V(s)| \leqslant I_1(s) + I_2(s)$$

where

$$\begin{aligned}
I_1(s) &= \left| \sum_{1 \leqslant i \leqslant n} (\varepsilon_i^2 - \sigma_i^2) \left(\sum_{1 \leqslant j \leqslant n} \text{sign}(X_j - X_i) Q(X_i, X_j, s) \right)^2 \right|, \\
I_2(s) &= \left| \sum_{1 \leqslant i \leqslant n} (\widehat{\sigma}_i^2 - \varepsilon_i^2) \left(\sum_{1 \leqslant j \leqslant n} \text{sign}(X_j - X_i) Q(X_i, X_j, s) \right)^2 \right|.
\end{aligned}$$

Consider $I_1(s)$. As in the proof of Lemma 12, let $u = u_n = n^{1/(4+\phi_1)}$ where $\phi_1 \in (-2, \phi)$. Let $\tilde{\varepsilon}_i = \varepsilon_i 1\{|\varepsilon_i| \leqslant u\}$ and $\tilde{\sigma}_i^2 = \text{E}[\tilde{\varepsilon}_i^2]$. It follows from Assumption A1 that $\text{P}(\max_{1 \leqslant i \leqslant n} |\tilde{\varepsilon}_i - \varepsilon_i| = 0) \to 1$, and $0 \leqslant \sigma_i^2 - \tilde{\sigma}_i^2 \lesssim 1/u^{2+\phi}$ uniformly over $i = \overline{1, n}$. Then $I_1(s) \lesssim I_{11}(s) + (nh)^3 h^{2k}/u^{2+\phi}$ w.p.a.1 where

$$I_{11}(s) = \left| \sum_{1 \leqslant i \leqslant n} (\tilde{\varepsilon}_i^2 - \tilde{\sigma}_i^2) \left(\sum_{1 \leqslant j \leqslant n} \text{sign}(X_j - X_i) Q(X_i, X_j, s) \right)^2 \right|.$$

Applying Bernstein inequality and using the union bound yields

$$\mathrm{P}(\max_{s\in\mathcal{S}_n} I_{11}(s)/V(s) > t) \leqslant 2\exp(\log p - C(nh_{\min})t^2/(1+u^2t)),$$

and so

$$\mathrm{P}(\max_{s\in\mathcal{S}_n} I_{11}(s)/V(s) > Cn^{-\kappa_3}) \to 0$$

for any $C > 0$ as long as conditions of the Lemma hold.

Consider $I_2(s)$. Clearly,

$$\begin{aligned} I_2(s) &\leqslant \sum_{1\leqslant i\leqslant n} ((\widehat{\sigma}_i - \varepsilon_i)^2 + 2|\varepsilon_i||\widehat{\sigma}_i - \varepsilon_i|) \left(\sum_{1\leqslant j\leqslant n} \mathrm{sign}(X_j - X_i)Q(X_i, X_j, s)\right)^2 \\ &\leqslant o_p(n^{-\kappa_1}) \sum_{1\leqslant i\leqslant n} (o_p(n^{-\kappa_1}) + |\varepsilon_i|) \left(\sum_{1\leqslant j\leqslant n} \mathrm{sign}(X_j - X_i)Q(X_i, X_j, s)\right)^2, \end{aligned}$$

and so $I_2(s)/V(s) = o_p(1)$ uniformly over $s \in \mathcal{S}_n$ by arguments similar to those used above. Combining presented results gives the asserted claim (c). □

Appendix E. Proofs for Section 6

Proof of Theorem 6. Denote $Y_i^0 = f(X_i) + \varepsilon_i$. Then $Y_i = Y_i^0 + Z_i^T\beta$ and $\tilde{Y}_i = Y_i^0 - Z_i^T(\widehat{\beta} - \beta)$. Therefore, $|\tilde{Y}_i - Y_i^0| \leqslant \|Z_i\|\|\widehat{\beta} - \beta\| = O_p(1/\sqrt{n})$ uniformly over $i = 1, ..., n$ and all models in $\mathcal{M}_{PL}$. So,

$$T = \max_{s\in\mathcal{S}_n} \sum_{1\leqslant i\leqslant n} \widehat{a}_i(s)\tilde{Y}_i = \max_{s\in\mathcal{S}_n} \sum_{1\leqslant i\leqslant n} \widehat{a}_i(s)Y_i^0 + o_p(1/\sqrt{\log p})$$

since

$$\begin{aligned} &\max_{s\in\mathcal{S}_n} \sum_{1\leqslant i\leqslant n} |\widehat{a}_i(s)(\tilde{Y}_i - Y_i^0)| = \max_{s\in\mathcal{S}_n} \sum_{1\leqslant i\leqslant n} |a_i(s)(\tilde{Y}_i - Y_i^0)|O_p(1) \\ &= \max_{s\in\mathcal{S}_n} \sum_{1\leqslant i\leqslant n} |a_i(s)|O_p(1/\sqrt{n})O_p(1) = o(\sqrt{n/\log p})O_p(1/\sqrt{n})O_p(1) = o_p(1/\sqrt{\log p}). \end{aligned}$$

The result follows by the argument similar to that used in the proof of Theorem 1. □

Proof of Theorem 7. The proof relies on the same notation as introduced in Section B of the Appendix with $f(x, z)$, $\bar{Q}(x_1, z_1, x_2, z_2, \bar{s})$, $\bar{s}$, $\bar{\mathcal{S}}_n$, and $\bar{p}$ substituting $f(x)$, $Q(x_1, x_2, s)$, s, $\mathcal{S}_n$, and p.

For $\mathcal{S} \subset \bar{\mathcal{S}}_n$ and $\eta \in (0, 1)$, let $c_\eta^{\mathcal{S},1}$ be the conditional η quantile of $\max_{\bar{s}\in\mathcal{S}} \sum_{1\leqslant i\leqslant n} a_i(\bar{s})\varepsilon_i\epsilon_i$ given $\{\varepsilon_i\}$. Since $\bar{A}_n(\log\bar{p})^{7/2} \to 0$, applying Corollary 6 (SC-d) of Chernozhukov, Chetverikov, and Kato (2012) shows that

$$\mathrm{P}(\max_{\bar{s}\in\mathcal{S}} \varepsilon(\bar{s}) \leqslant c_\eta^{\mathcal{S},1}) = \eta + o(1)$$

uniformly over all $\mathcal{S} \subset \bar{\mathcal{S}}_n$ and $\eta \in (0,1)$, and so $\max_{\bar{s} \in \bar{\mathcal{S}}_n} |\varepsilon(\bar{s})| = O_p(\sqrt{\log \bar{p}})$. In addition, the result of Lemma 10 holds under the conditions (i)-(iv) of Assumption A12, and so

$$\mathrm{P}(\max_{\bar{s} \in \mathcal{S}} \varepsilon(\bar{s}) \leqslant c_\eta^{\mathcal{S},0}) = \eta + o(1)$$

uniformly over all $\mathcal{S} \subset \bar{\mathcal{S}}_n$ and $\eta \in (0,1)$, which gives the result analogous to that in Lemma 8. Further,

$$T = \max_{\bar{s} \in \bar{\mathcal{S}}_n} \sum_{1 \leqslant i \leqslant n} \widehat{a}_i(\bar{s}) Y_i = \max_{(s,z) \in \bar{\mathcal{S}}_n} \sum_{1 \leqslant i \leqslant n} \widehat{a}_i(\bar{s})(f(X_i, z) + \varepsilon_i) + o_p(1/\sqrt{\log \bar{p}})$$

by conditions (v) and (vi) of Assumption A12. Therefore, the result follows by the argument similar to that used in the proof of Theorem 1.

□

References

Anderson, E., and D. Schmittlein (1984): "Integration of the Sales Force: An Empirical Examination," *RAND Journal of Economics*, 15, 385–395.

Andrews, D. W. K., and X. Shi (2010): "Inference Based on Conditional Moment Inequalities," *Cowles Foundation Discussion Paper, No 1761.*

Angeletos, M., and I. Werning (2006): "Information Aggregation, Multiplicity, and Volatility," *American Economic Review*, 96(5).

Athey, S. (2002): "Monotone Comparative Statics under Uncertainty," *The Quarterly Journal of Economics*, 117, 187–223.

Beraud, Y., S. Huet, and 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.

Boucheron, S., O. Bousquet, and G. Lugosi (2004): "Concentration Inequalities," *Advanced Lectures in Machine Learning*, pp. 208–240.

Bowman, A. W., M. C. Jones, and I. Gijbels (1998): "Testing Monotonicity of Regression," *Journal of Computational and Graphical Statistics*, 7, 489–500.

Chatterjee, S. (2005): "A Simple Invariance Theorem," *arXiv:math/0508213v1.*

Chernozhukov, V., D. Chetverikov, and K. Kato (2011): "Anti-Concentration and Honest Adaptive Confidence Bands," *working paper*, pp. 1–20.

——— (2012): "Multiplier and Gaussian Comparison Theorems for High-Dimensional Inference," *working paper*, pp. 1–30.

Chernozhukov, V., S. Lee, and A. M. Rosen (2009): "Intersection Bounds: Estimation and Inference," *CEMMAP working paper CWP 19/09.*

Chetverikov, D. (2012): "Adaptive Test of Conditional Moment Inequalities," *arXiv:1201.0167v2.*

Delgado, M., and J. Escanciano (2010): "Testing Conditional Monotonicity in the Absence of Smoothness," *working paper*, pp. 1–18.

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

Dumbgen, L., and V. G. 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., and S. F. Ellison (2011): "Strategic Entry Deterrence and the Behavior of Pharmaceutical Incumbents Prior to Patent Expiration," *American Economic Journal: Microeconomics*, 3, 1–36.

Fan, J., and Q. Yao (1998): "Efficient Estimation of Conditional Variance Functions in Stochastic Regression," *Biometrica*, 85, 645–660.

Ghosal, S., A. Sen, and A. van der Vaart (2000): "Testing Monotonicity of Regression," *The Annals of Statistics*, 28, 1054–1082.

Gijbels, I., P. Hall, M. C. Jones, and I. Koch (2000): "Tests for Monotonicity of a Regression Mean with Guaranteed Level," *Biometrica*, 87, 663–673.

Hall, P., and N. Heckman (2000): "Testing for Monotonicity of a Regression Mean by Calibrating for Linear Functions," *The Annals of Statistics*, 28, 20–39.

Hardle, W., and E. Mammen (1993): "Comparing Nonparametric Versus Parametric Regression Fits," *The Annals of Statistics*, 21, 1926–1947.

Hardle, W., and A. Tsybakov (1997): "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.

Holmstrom, B., and P. Milgrom (1994): "The Firm as an Incentive System," *The American Economic Review*, 84, 972–991.

Horowitz, J. L. (2009): *Semiparametric and Nonparametric Methods in Econometrics*. Springer Series in Statistics.

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

Lee, S., O. Linton, and Y. Whang (2009): "Testing for Stochastic Monotonicity," *Econometrica*, 77(2), 585–602.

Lee, S., K. Song, and Y. J. Whang (2011): "Testing function inequalities," *CEMMAP working paper CWP 12/11*.

Lehmann, E., and J. Romano (2005): *Testing Statistical Hypotheses*. Springer.

Liu, R. (1988): "Bootstrap Procedures under niid Models," *The Annals of Statistics*, 16(4), 1696–1708.

Mammen, E. (1993): "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *The Annals of Statistics*, 21(1), 255–285.

Manski, C. F., and V. Pepper, J. (2000): "Monotone Instrumental Variables: With an Application to Returns to Schooling," *Econometrica*, 68, 997–1010.

Merton, R. (1974): "On the Pricing of Corporate Debt: the Risk Structure of Interest Rates," *Journal of Finance*, 29, 449–470.

Milgrom, P., and C. Shannon (1994): "Monotone Comparative Statics," *Econometrica*, 62, 157–180.

Morris, S., and S. Shin, H. (1998): "Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks," *The American Economic Review*, 88(3), 587–597.

——— (2001): "Global Games: Theory and Application," *Cowles Foundation Discussion Paper, No 1275R*, pp. 1–70.

——— (2004): "Coordination Risk and the Price of Debt," *European Economic Review*, 48, 133–153.

Muller, H.-G. (1991): "Smooth Optimum Kernel Estimators near Endpoints," *Biometrica*, 78(3), 521–30.

Muller, H. G., and U. Stadtmuller (1987): "Estimation of Heteroscedasticity in Regression Analysis," *The Annals of Statistics*, 15, 610–625.

Newey, W. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

Pollard, D. (1984): *Convergence of Stochastic Processes*. Springer-Verlag.

Poppo, L., and T. Zenger (1998): "Testing Alternative Theories of the Firm: Transaction Cost, Knowlegde-Based, and Measurement Explanations of Make-or-Buy Decisions in Information Services," *Strategic Management Journal*, 19(9), 853–877.

Rice, J. (1984): "Bandwidth Choice for Nonparametric Kernel Regression," *The Annals of Statistics*, 12, 1215–1230.

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

Romano, J., and M. Wolf (2005a): "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 73(4), 1237–1282.

Romano, J., P., and A. M. Shaikh (2010): "Inference for the Identified Sets in Partially Identified Econometric Models," *Econometrica*, 78, 169–211.

Romano, J., P., and M. Wolf (2005b): "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 100, 94–108.

Schlee, W. (1982): "Nonparametric Tests of the Monotony and Convexity of Regression," in *Nonparametric Statistical Inference*. Amsterdam: North-Holland.

Tirole, J. (1988): *The Theory of Industrial Organization*. Cambridge, MA: MIT Press.

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

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

Wang, J., and 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(4), 1261–1295.

Supplementary Appendix

This supplementary Appendix contains additional simulation results. In particular, I consider the test developed in this paper with weighting functions of the form given in equation (2) with $k = 1$. The simulation design is the same as in Section 7. The results are presented in table 2. For ease of comparison, I also repeat the results for the tests of GSV, GHJK, and HH in this table. Overall, the simulation results in table 2 are similar to those in table 1, which confirms the robustness of the findings in this paper.

TABLE 3. Results of Monte Carlo Experiments

Noise	Case	Sample	Proportion of Rejections for								
			GSV	GHJK	HH	CS-PI	CS-OS	CS-SD	IS-PI	IS-OS	IS-SD
normal	1	100	.118	.078	.123	.129	.129	.129	.166	.166	.166
		200	.091	.051	.108	.120	.120	.120	.144	.144	.144
		500	.086	.078	.105	.121	.121	.121	.134	.134	.134
normal	2	100	0	.001	0	.002	.009	.009	.006	.024	.024
		200	0	.002	0	.001	.012	.012	.007	.016	.016
		500	0	.001	0	.002	.005	.005	.005	.016	.016
normal	3	100	0	.148	.033	.238	.423	.432	0	0	0
		200	.010	.284	.169	.639	.846	.851	.274	.615	.626
		500	.841	.654	.947	.977	.995	.996	.966	.994	.994
normal	4	100	.037	.084	.135	.159	.228	.231	.020	.040	.040
		200	.254	.133	.347	.384	.513	.515	.372	.507	.514
		500	.810	.290	.789	.785	.833	.833	.782	.835	.836
uniform	1	100	.109	.079	.121	.120	.120	.120	.200	.200	.200
		200	.097	.063	.109	.111	.111	.111	.154	.154	.154
		500	.077	.084	.107	.102	.102	.102	.125	.125	.125
uniform	2	100	.001	.001	0	0	.006	.006	.015	.031	.031
		200	0	0	0	.001	.009	.009	.013	.021	.024
		500	0	.003	0	.003	.012	.012	.011	.021	.021
uniform	3	100	0	.151	.038	.225	.423	.433	0	0	0
		200	.009	.233	.140	.606	.802	.823	.261	.575	.590
		500	.811	.582	.947	.976	.993	.994	.971	.990	.991
uniform	4	100	.034	.084	.137	.150	.216	.219	.020	.046	.046
		200	.197	.116	.326	.355	.483	.488	.328	.466	.472
		500	.803	.265	.789	.803	.852	.855	.796	.859	.861

Nominal Size is 0.1. GSV, GHJK, and HH stand for the tests of Ghosal, Sen, and van der Vaart (2000), Gijbels, Hall, Jones, and Koch (2000), and Hall and Heckman (2000) respectively. CS-PI, CS-OS, and CS-SD refer to the test developed in this paper with σ_i estimated using Rice's formula and plug-in, one-step, and stepdown critical values respectively. Finally, IS-PI, IS-OS, and IS-SD refer to the test developed in this paper with σ_i estimated by $\widehat{\sigma}_i = \widehat{\varepsilon}_i$ and plug-in, one-step, and stepdown critical values respectively.