New Methods for Econometric Inference

by

Denis Chetverikov

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013

Author ....
Department of Economics
May 15, 2013

Certified by....................
Victor Chernozhukov
Professor
Thesis Supervisor

Certified by........ ....
Anna Mikusheva
Castle-Krob Career Development Associate Professor
Thesis Supervisor

Accepted by.. ....
Michael Greenstone
Chairman, Department Committee on Graduate Theses

New Methods for Econometric Inference

by

Denis Chetverikov

Submitted to the Department of Economics
on May 15, 2013, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

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. Chapter 1 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. Chapter 1 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.

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

In chapter 3 (coauthored with Victor Chernozhukov and Kengo Kato), we derive a central limit theorem for the maximum of a sum of high dimensional random vectors. Specifically, we establish conditions under which the distribution of the maximum is approximated by that of the maximum of a sum of the Gaussian random vectors with the same covariance matrices as the original vectors. The key innovation of this result is that it applies even when the dimension of random vectors (p) is large compared to the sample size (n); in fact, p can be much larger than n. We also show that the distribution of the maximum of a sum of the random vectors with unknown covariance matrices can be consistently estimated by the distribution of the maximum of a sum of the conditional Gaussian random vectors obtained by multiplying the original vectors with i.i.d. Gaussian multipliers. This is the multiplier bootstrap procedure. Here too, p can be large or even much larger than n. These distributional approximations, either Gaussian or conditional Gaussian, yield a high-quality approximation to the distribution of the original maximum, often with approximation error decreasing polynomially in the sample size, and hence are of interest in many applications. We demonstrate how our central limit theorem and the multiplier bootstrap can be used for high dimensional estimation, multiple hypothesis testing, and adaptive specification testing. All these results contain non-asymptotic bounds on approximation errors.

Thesis Supervisor: Victor Chernozhukov
Title: Professor

Thesis Supervisor: Anna Mikusheva
Title: Castle-Krob Career Development Associate Professor

Acknowledgments

I owe special thanks to my advisors, Victor Chernozhukov and Anna Mikusheva for their time and efforts in explaining me beauties of the world of Econometrics. Victor served as a role model for me. His enthusiasm and energy were very important for me. Every time I left Victor's office, I felt I could work three days and three nights without a sleep. I thank Anna for introducing me into the field of Econometrics. I am also greatly indepted to Whitney Newey for his wise advices and Kengo Kato working with whom has been a lot of fun and from whom I learned a lot.

Most of all I am grateful to my wife for her love and support. She helped me a lot during the hard times I got stuck with deriving new methods and proving theorems. She definitely sacrificed more than I did to complete this Ph.D.

Contents

1 Testing Regression Monotonicity in Econometric Models 15

1.1 Introduction 15

1.2 Motivating Examples 19

1.3 The Test 22

1.3.1 The General Test Statistic 22

1.3.2 Typical Weighting Functions 24

1.3.3 Comparison with Other Known Tests 27

1.3.4 Estimating σ_i 29

1.3.5 Simulating the Critical Value 30

1.4 Theory under High-Level Conditions 33

1.5 Verification of High-Level Conditions 40

1.6 Models with Multivariate Covariates 42

1.6.1 Partially Linear Model 42

1.6.2 Nonparametric Model 44

1.7 Monte Carlo Simulations 47

1.8 Empirical Application 52

1.9 Conclusion 55

1.10 Appendix A. Implementation Details 56

1.10.1 Plug-in Approach 57

1.10.2 One-Step Approach 57

1.10.3 Stepdown Approach 57

1.11 Appendix B. Additional Notation 58

1.12 Appendix C. Proofs for section 1.4 59

1.12.1 Auxiliary Lemmas 59

1.12.2 Proofs of Theorems 67

1.13 Appendix D. Proofs for Section 1.5 71

1.14 Appendix E. Proofs for Section 1.6 76

2 Adaptive Test of Conditional Moment Inequalities 79

2.1 Introduction 79

2.2 Examples 85

2.3 Test 87

2.3.1 The Test Statistic 88

2.3.2 Critical Values 90

2.3.3 Estimating Σ_i 91

2.4 Main Results 93

2.4.1 Assumptions 93

2.4.2 Size Properties of the Test 97

2.4.3 Consistency Against a Fixed Alternative 98

2.4.4 Consistency Against Local One-Dimensional Alternatives . . . 99

2.4.5 Uniform Consistency Against Holder Smoothness Classes . . . 99

2.4.6 Lower Bound on the Minimax Rate of Testing 100

2.5 Extentions 102

2.6 Monte Carlo Results 105

2.7 Conclusions 110

2.8 Appendix. Proofs 111

2.8.1 Lemma on the Estimator of Σ_i 111

2.8.2 Continuity of the Square Root Operator on the Set of Positive Semidefinite Matrices 114

2.8.3 Primitive Conditions for A1 115

2.8.4 Anticoncentration Inequality for the Maximum of Gaussian Random Variables 117

2.8.5 Result on Gaussian Random Variables 118
2.8.6 Preliminary Technical Results 118
2.8.7 Proofs of Theorems 127

3 Central Limit Theorems and Multiplier Bootstrap when p is much larger than n 135

3.1 Introduction 135
3.1.1 Organization of the paper 140
3.1.2 Notation 141
3.2 Central Limit Theorems for Maxima of Non-Gaussian Sums 141
3.2.1 Comparison Theorems and Non-Asymptotic Gaussian Approximations 141
3.2.2 Examples of Applications 146
3.3 Multiplier Bootstrap 148
3.3.1 A Gaussian-to-Gaussian Comparison Theorem 148
3.3.2 Multiplier Bootstrap Theorems 149
3.3.3 Examples of Applications: Revisited 152
3.4 Application: Dantzig Selector in the Non-Gaussian Model 152
3.4.1 Homoscedastic case 153
3.4.2 Heteroscedastic case. 157
3.4.3 Some Extensions 159
3.5 Appendix A. Preliminaries 160
3.5.1 A Useful Maximal Inequality 160
3.5.2 Properties of the Smooth Max Function 160
3.5.3 Lemma on Truncation 162
3.6 Appendix B. Proofs for Section 2 163
3.6.1 Proof of Theorem 13 163
3.6.2 Proof of Corollary 3 165
3.6.3 Proof of Theorem 14 166
3.6.4 Proof of Corollary 4 171

3.6.5 Proof of Lemma 33 171
3.6.6 Proof of Corollary 5 171
3.7 Appendix C. Proofs for Section 3.3 172
3.7.1 Proof of Lemma 35 172
3.7.2 Proof of Lemma 36 173
3.7.3 Proof of Theorem 15 173
3.7.4 Proof of Theorem 16 173
3.7.5 Proof of Corollary 6 174
3.8 Appendix D. Proofs for Section 3.4 175
3.8.1 Proof of Theorem 17 175
3.8.2 Proof of Theorem 18 176
3.9 Appendix E. A note on relation between Slepian and Stein type methods for normal approximations 178
3.10 Appendix F. Additional proofs 180
3.10.1 Proof of Lemma 43 180
3.10.2 Proof of Corollary 4 181
3.11 Appendix G. Application: Multiple Hypothesis Testing via the Step-down Method 184
3.12 Appendix H. Application: Adaptive Specification Testing 188
3.13 Appendix I. Proofs for Section 3.11 190
3.13.1 Proof of Theorem 19 190
3.14 Appendix J. Proofs for Section 3.12 195
3.14.1 Proof of Theorem 20 195

List of Figures

1-1 Regression Functions Illustrating Different Deviations from $\mathcal{H}_0$ 26

1-2 Regression Functions Used in Simulations 49

2-1 The difference between the rejection probabilities of the test developed in this paper and of the test of Andrews and Shi (2013) (with RMS and GMS critical values correspondingly). The nominal size is 10%. Results are based on 500 simulations. The figure shows that the rejection probability of the test developed in this paper is higher than that of the test of Andrews and Shi (2013) in most cases and is strictly higher over a wide region of parameter values. 110

3-1 P-P plots comparing distributions of T_0 and Z_0 in the example motivated by the problem of seleting the penalty level of the Dantzig selector. Here x_{ij} are generated as $x_{ij} = z_{ij}\varepsilon_i$ with $\varepsilon_i \sim t(4)$, (a t-distribution with four degrees of freedom), and z_{ij} are non-stochastic (simulated once using $U[0,1]$ distribution independently across i and j). The dashed line is 45°. The distributions of T_0 and Z_0 are close, as (qualitatively) predicted by the CLT derived in the paper: see Corollaries 3 or 4. The quality of the Gaussian approximation is particularly good for the tail probabilities, which is most relevant for practical applications. 137

List of Tables

1.1 Results of Monte Carlo Experiments 51

1.2 Incumbent Behavior versus Market Size: Monotonicity Test p-value . 55

1.3 Results of Monte Carlo Experiments 78

2.1 Results of Monte Carlo Experiments, $n = 250$ 108

2.2 Results of Monte Carlo Experiments, $n = 500$ 109

Chapter 1

Testing Regression Monotonicity in Econometric Models

1.1 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 [82] and [13]. 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 [44], these tests can also be used to provide evidence of 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

$$Y_i = f(X_i) + \varepsilon_i,\ i = 1, 2, 3, ... \tag{1.1}$$

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 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 [36], [5], and [37], 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 [103] and [102]. 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 [50] and [49] (from now on, GHJK and GSV, respectively) are based on the signs of $(Y_{i+k} - Y_i)(X_{i+k} - X_i)$. [57] (from now on, HH) developed a test based on the slopes of local linear estimates of f. The list of other papers includes [105], [21], [42], [43], [18], and [112]. In a contemporaneous work, [73] derive another approach to testing monotonicity based on L_p-functionals. An advantage of their

[1] The exception is [112] 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.

method is that the asymptotic distribution of their test statistic in the least favorable model under $\mathcal{H}_0$ turns out to be $N(0,1)$. A disadvantage of their method, however, is that their test is not adaptive. [72] and [40] 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, [109]), the empirical evidence is rather limited. [44] 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 1.2 provides motivating examples. Section 1.3 describes the general test statistic and gives several methods to simulate the critical value. Section 1.4 contains the main results under high-level conditions. Section 1.5 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 1.6 studies the cases of partially linear and fully nonparametric models with multiple covariates. Section 1.7 presents a small Monte Carlo simulation study. Section 1.8 describes the empirical application. Section 1.9 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.

1.2 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 [61] 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 [2] 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 [2] and [95].

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, [84], [85], and [7].

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 1.8. Below I provide another example concerned with the problem of debt pricing. This example is based on [85]. Consider a model where investors hold a collateralized debt. The debt will yield a fixed payment (1) in the future if it is rolled over and an underlying project is successful. Otherwise the debt will yield nothing (0). Alternatively, all investors have an option of not rolling over and getting the value of the collateral, $\kappa \in (0,1)$, immediately. The probability that the project turns out to be successful depends on the fundamentals, θ, and on how many investors roll over. Specifically, assume that the project is successful if θ exceeds the proportion of investors who roll over. Under global game reasoning, if private information possessed by investors is sufficiently accurate, the project will succeed if and only if $\theta \geqslant \kappa$; see [85] for details. Then ex ante value of the debt is given by

$$V(\kappa) = \kappa \cdot \mathrm{P}(\theta < \kappa) + 1 \cdot \mathrm{P}(\theta \geqslant \kappa)$$

and the derivative of the ex ante debt value with respect to the collateral value is

$$\frac{dV(\kappa)}{d\kappa} = \mathrm{P}(\theta < \kappa) - (1-\kappa)\frac{d\mathrm{P}(\theta < \kappa)}{d\kappa}$$

The first and second terms on the right hand side of the equation above represent direct and strategic effects correspondingly. The strategic effect represents coordination failure among investors. It arises because high value of the collateral leads investors to believe that many other investors will not roll over, and the project will not be successful even though the project is profitable. [86] argue that this effect is important for understanding anomalies in empirical implementation of the standard debt pricing theory of [81]. A natural question is how to prove existence of this effect in the data. Note that in the absense of strategic effect, the relation between value of the debt and value of the collateral will be monotonically increasing. If strategic effect is sufficiently strong, however, it can cause non-monotonicity in this relation. Therefore, one can detect the existence of the strategic effect and coordination failure by testing 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 and coordination failure.

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 $E[y_i(t)]$. [79] 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, [82].

1.3 The Test

1.3.1 The General Test Statistic

Recall that I consider a model given in equation (1.1), 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 1.4 and 1.5, I assume that $X_i \in \mathbb{R}$. The case where $X_i \in \mathbb{R}^d$ for $d > 1$ is considered in Section 1.6.

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 and there exist a pair (i, j) such that $X_i < X_j$ and $f(X_i) > f(X_j)$, then 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 [63], 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's 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 $\mathcal{S}_n$.

1.3.2 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

$$Q(x_1, x_2, (x, h)) = |x_1 - x_2|^k K\left(\frac{x_1 - x}{h}\right) K\left(\frac{x_2 - x}{h}\right) \qquad (1.2)$$

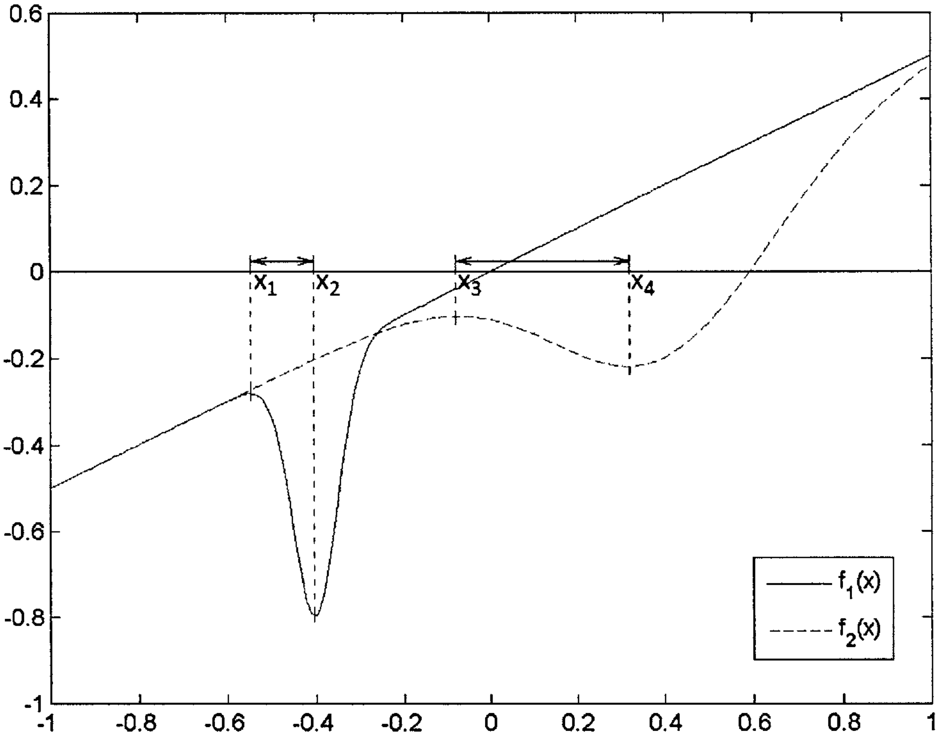
[2] The kernel function is called positive if it is positive on its support.

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 higher signal-to-noise ratio. So, given that $[x_3, x_4]$ is longer than $[x_1, x_2]$, the optimal pair (x, h) to test against f_2 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 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.

Figure 1-1: Regression Functions Illustrating Different Deviations from $\mathcal{H}_0$



Another interesting choice of the weighting functions is

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

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 also recommend setting $u=0.5$, $h_{\min}=0.4h_{\max}(\log n/n)^{1/3}$, and $k=0$ or 1. I refer to this $\mathcal{S}_n$ as a basic set of weighting functions. This choice of parameters is consistent with the theory presented in sections 1.4 and 1.5 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 $n=100$ and the sequence $\{X_i\}$ is distributed uniformly.

1.3.3 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:

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

whereas setting $k=0$ in equation (1.2) yields

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

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

$$b(s) = \frac{\sum_{s_1 < i \leqslant s_2} Y_i \sum_{s_1 < j \leqslant s_2} (X_i - X_j)}{(s_2 - s_1) \sum_{s_1 < i \leqslant s_2} X_i^2 - (\sum_{s_1 < i \leqslant s_2} X_i)^2}, \tag{1.3}$$

where $s = (s_1, s_2)$. Note that the denominator of (1.3) is nonstochastic, and so it disappears after studentization. In addition, simple rearrangements show that the numerator in (1.3) is up to the sign is equal to

$$(1/2) \sum_{1 \leqslant i,j \leqslant n} (Y_i - Y_j)(X_j - X_i) 1\{x - h \leqslant X_i \leqslant x + h\} 1\{x - h \leqslant X_j \leqslant x + h\} \tag{1.4}$$

for some x and h. On the other hand, setting $k = 1$ in equation (1.2) yields

$$b(s) = (1/2) \sum_{1 \leqslant i,j \leqslant n} (Y_i - Y_j)(X_j - X_i) K\left(\frac{X_i - x}{h}\right) K\left(\frac{X_j - x}{h}\right). \tag{1.5}$$

Noting that expression in (1.4) is proportional to that on the right hand side in (1.5) with $K(\cdot) = 1\{[-1, +1]\}(\cdot)$ implies that the HH test statistic is a special case of those studied in this paper.

1.3.4 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 1.4, 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 [90], theorem 1) or local polynomial regression (see [110], 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, [87]. 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 $\widehat{\sigma}_i - \sigma_i = o_p(n^{-\kappa_2})$ uniformly over $i = \overline{1,n}$ for some $\kappa_2 > 0$. 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 [97], given by

$$\widehat{\sigma} = \left(\frac{1}{2n} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2 \right)^{1/2}, \tag{1.6}$$

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

$$\widehat{\sigma}_i = \left(\frac{1}{2|J(i)|} \sum_{j \in J(i): j+1 \in J(i)} (Y_{j+1} - Y_j)^2 \right)^{1/2}. \tag{1.7}$$

I refer to (1.7) 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 1.5 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 [88], [46], [63], [59], and [25].

1.3.5 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 [113] and used, among many others, by [77], [78], [58], [63], and [37]. See also [33]. The three methods are arranged in terms of increasing power and computational complexity. The validity of all three methods is established in theorem 8. 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 level α. The plug-in approach is based on two observations. First, under $\mathcal{H}_0$,

$$\begin{aligned} b(s) &= (1/2) \sum_{1 \leqslant i,j \leqslant n} (f(X_i) - f(X_j) + \varepsilon_i - \varepsilon_j) \operatorname{sign}(X_j - X_i) Q(X_i, X_j, s) \quad (1.8) \\ &\leqslant (1/2) \sum_{1 \leqslant i,j \leqslant n} (\varepsilon_i - \varepsilon_j) \operatorname{sign}(X_j - X_i) Q(X_i, X_j, s) \quad (1.9) \end{aligned}$$

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 1.10 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 $c_{1-\alpha}^{PI}$. Therefore, $c_{1-\alpha}^{PI}$ 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. In this paper, I develop two selection procedures that 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

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

Then the one-step critical value $c^{OS}_{1-\alpha}$ is the conditional $(1-\alpha)$ quantile of the simulated statistic $T^\star = T(\{X_i, Y_i^\star\}, \{\widehat{\sigma}_i\}, \mathcal{S}_n^{OS})$ given $\{\widehat{\sigma}_i\}$ and $\mathcal{S}_n^{OS}$ where $Y_i^\star = \widehat{\sigma}_i \epsilon_i$ for $i = \overline{1,n}$.[3] Intuitively, the one-step critical value works because the weighting functions corresponding to elements of the set $\mathcal{S}_n \backslash \mathcal{S}_n^{OS}$ 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

$$(1/2) \sum_{1 \leqslant i,j \leqslant n} (f(X_i) - f(X_j)) \text{sign}(X_j - X_i) Q(X_i, X_j, s)/(\widehat{V}(s))^{1/2} > -c^{PI}_{1-\gamma_n} \quad (1.10)$$

belongs to the set $\mathcal{S}_n \backslash \mathcal{S}_n^{OS}$ 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 (1.10) 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 $\mathcal{S}_n^{OS}$. 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

[3] If $\mathcal{S}_n^{OS}$ turns out to be empty, assume that $\mathcal{S}_n^{OS}$ consists of one randomly chosen element of $\mathcal{S}_n$.

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, [60], [100], [103], and [102], and [75] 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 $T^\star = T(\{X_i, Y^\star_i\}, \{\widehat{\sigma}_i\}, \mathcal{S}^l_n)$ given $\{\widehat{\sigma}_i\}$ and $\mathcal{S}^l_n$ where $Y^\star_i = \widehat{\sigma}_i \epsilon_i$ 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 $T^\star = T(\{X_i, Y^\star_i\}, \{\widehat{\sigma}_i\}, \mathcal{S}^{SD}_n)$ given $\{\widehat{\sigma}_i\}$ and $\mathcal{S}^{SD}_n$ where $Y^\star_i = \widehat{\sigma}_i \epsilon_i$ 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.

1.4 Theory under High-Level Conditions

This section describes the high-level assumptions used in the 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 1.3.4. 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 1.3.4 for different available estimators. See also Lemma 2 in Section 1.5 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 $(V(s)/\widehat{V}(s))^{1/2} - 1 = o_p(n^{-\kappa_3})$ 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 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.

A 5. *$nA_n^4(\log p)^7 = o(1)$ where $p = |\mathcal{S}_n|$, the number of elements in the set $\mathcal{S}_n$. In addition, if A2 holds, then $\log p/n^{(1/4)\wedge\kappa_1\wedge\kappa_3} = o(1)$, and 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 (1.1), 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*

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

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

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

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 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 [33] and [34]. 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.

[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.

(iv) To obtain a particular application of the general result presented in this theorem, consider the basic set of weighting functions introduced in subsection 1.3.2. Then the number of weighting functions in the set $\mathcal{S}_n$ is bounded from above by some polynomial in n, and so $\log p \leqslant C \log n$. Lemma 3 in the next section then implies that Assumptions A4 and A5 hold (under mild conditions on $K(\cdot)$ stated in Lemma 3), and so the result of Theorem 8 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.

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 singleton w.p.a.1, which means that $\mathrm{P}\{c^{OS}_{1-\alpha} \leqslant C\} \to 1$ and $\mathrm{P}\{c^{SD}_{1-\alpha} \leqslant C\} \to 1$ for some $C > 0$. On the other hand, $\mathrm{P}(c^{PI}_{1-\alpha} > C) \to 1$ for the same C 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 [41]) 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 [41] 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, [5], [74], and the discussion in [63].

(ii) Suppose that $\mathcal{S}_n$ consists of the basic set of weighting functions. Then $\log p \leqslant C \log n$*, 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 the 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 the 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.

A8. *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.

A9. *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. 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. The following theorem gives the uniform rate of consistency.

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*

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

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 [63]. 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 S_n consists of the basic set of weighting functions. Then Assumption A9 holds. In addition, Lemma 1 gives conditions that suffice for Assumption A8, and Lemma 3 shows that Assumptions A4 and A5 are satisfied under mild conditions on $K(\cdot)$. So, Theorem 4 implies that the test with this 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 uniformly consistent against models with $\inf_{x \in [s_l, s_r]} f_n^{(1)}(x) > -C(\log n/n)^{\beta/(2\beta+3)}$ for some sufficiently small $C > 0$ if it controls size. Therefore, the test based on the basic set of weighting functions 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 small 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 the basic set of weighting functions 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.*

1.5 Verification of High-Level Conditions

This section provides conditions that are sufficient for the assumptions used in Section 1.4. 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 1.4 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 (1.7). Suppose also that (i) Assumption A1 holds, (ii)* $\log n = o(n^{-2\kappa_2'+\phi/(4+\phi)}b_n)$ *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)* $|f(X_i) - f(X_j)| \leqslant C|X_i - X_j|$ *uniformly over* $i,j = \overline{1,n}$*, and (v)* $|\sigma_i^2 - \sigma_j^2| \leqslant C|X_i - X_j|$ *uniformly over* $i,j = \overline{1,n}$*. Then* $\max_{1\leqslant i\leqslant n}|\widehat{\sigma}_i - \sigma_i| = O_p(b_n + (nb_n)^{-1/2} + n^{-\kappa_2'})$.

Note Assumption (iii) of this lemma follows from A8 whenever $b_n \geqslant Cn^{-1/3}$ for sufficiently large constant $C > 0$, 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 whenever $b_n + (nb_n)^{-1/2} \leqslant Cn^{-c}$ and $\log n \leqslant Cn^{\phi/(4+\phi)-c}b_n$ for some constants $c, C > 0$.

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 Assump-*

[5] Recall that in section 1.4, 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 $\{X_i\}_{1\leqslant i\leqslant\infty}$.

tion A4 holds with $\kappa_3 = \kappa_2$; (c) if Assumption A2 is satisfied, then Assumption A4 holds with any $\kappa_3 \leqslant \kappa_1$ as long as $\log p = o(h_{\min} n^{1-2\kappa_3})$ and $\log p = o(h_{\min} n^{1/2-\kappa_3})$.

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.

1.6 Models with Multivariate Covariates

Most empirical studies contain additional covariates that should be controlled for. In this section, I extend the results presented in Section 1.4 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. The power properties of the test can be obtained using the arguments closely related to those used in Theorems 2, 3, and 4.

1.6.1 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

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

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 1.3 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 [99], 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 [62] 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 estimators $\widehat{\sigma}_i$ of $\sigma_i = (\mathrm{E}[\varepsilon_i^2])^{1/2}$ satisfy either $\widehat{\sigma}_i = \widehat{\varepsilon}_i = Y_i - \widehat{f}(X_i) - Z_i^T\widehat{\beta}$ (here $\widehat{f}(X_i)$ is some estimator of $f(X_i)$, which is uniformly consistent over $i = \overline{1, n}$) or $\widehat{\sigma}_i$ is some uniformly consistent estimator of σ_i. The critical value for the test is simulated by one of the methods (plug-in, one-step, or stepdown) described in Section 1.3 using the data $\{X_i, \tilde{Y}_i\}$, estimators $\{\widehat{\sigma}_i\}$, and the set of weighting functions $\mathcal{S}_n$. As in Section 1.3, 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.

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)* $\|\widehat{\beta} - \beta\| = O_p(n^{-1/2})$, *and (iii) uniformly over all* $s \in \mathcal{S}_n$, $\sum_{1 \leqslant i,j \leqslant n} Q(X_i, X_j, s)/V(s)^{1/2} = o(\sqrt{n/\log p})$.

Let $\mathcal{M}_{PL}$ denote any set of models in $\mathcal{M}$ such that A11 is 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*

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

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

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

1.6.2 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 1.3. 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

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

$\bar{K} : \mathbb{R}^d \to \mathbb{R}$ 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. It is important here that the auxiliary bandwidth value $\bar{h}(\bar{s})$ depends on $\bar{s}$. For example, if kernel weighting functions are used, so that $\bar{s} = (x, h, z)$, then one has to choose $\bar{h} = \bar{h}(\bar{s})$ so that $nh\bar{h}^d \to \infty$ and $nh\bar{h}^{d+2} \to 0$ polynomially fast uniformly over $\bar{s} \in \bar{\mathcal{S}}_n$; see discussion after the statement of Assumption A12. 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 1.3 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 $c^{PI}_{1-\alpha}$, $c^{OS}_{1-\alpha}$, and $c^{SD}_{1-\alpha}$ denote the plug-in, one-step, and stepdown critical values correspondingly. In addition, let

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

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.

A 12. *(i)* $\mathrm{P}(|\varepsilon_i| \geqslant u) \leqslant \exp(-u/C_{10})$ *for all* $u \geqslant 0$ *and* $\sigma_i \geqslant C_2$ *for all* $i = \overline{1,n}$*, (ii)* $\bar{A}_n(\log(\bar{p}n))^{7/2} = o(1)$*, (iii)* $\bar{h}(\bar{s})\sum_{1\leqslant i,j\leqslant n}\bar{Q}(X_i, Z_i, X_j, Z_j, \bar{s})/(V(\bar{s}))^{1/2} = o(1/\sqrt{\log \bar{p}})$ *uniformly over* $\bar{s}\in\bar{\mathcal{S}}_n$*, and (iv) the regression function* f *has uniformly bounded first order partial derivatives.*

Condition (i) of this assumption imposes that ε_i's have sub-exponential tails, which is stronger than Assumption A1. It holds, for example, if ε_i's have normal distribution. Condition (iv) is a smoothness assumption. Conditions (ii) and (iii) 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 $\log \bar{p} \leqslant C \log n$. 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 typically $\bar{A}_n \leqslant C \max_{(h,\bar{h})\in\bar{\mathcal{S}}_{n,h}} 1/(nh\bar{h}^d)^{1/2}$ and $\sum_{1\leqslant i,j\leqslant n}\bar{Q}(X_i, Z_i, X_j, Z_j, \bar{s})/(V(\bar{s}))^{1/2} \leqslant C(nh\bar{h}^d)^{1/2}$. Therefore, conditions (ii) and (iii) 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 1.4 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 a bias term in the test statistic. Conditions (iii) and (iv) of Assumption A12 ensure that this bias is asymptotically negligible relative to the concentration rate of the test statistic. The difficulty, however, is that condition (iii) is inconsistent with $n\bar{A}_n^4(\log \bar{p})^7 \to 0$ imposed in Assumption A5 (where I replaced A_n and p by their multivariate analogs $\bar{A}_n$ and $\bar{p}$). Indeed, condition $n\bar{A}_n^4(\log \bar{p})^7 \to 0$ requires $nh^2\bar{h}^{2d} \to \infty$, and so it contradicts to $nh\bar{h}^{d+2} \to 0$ (if $d \geqslant 2$), which follows from condition (iii) of A12. To deal with this problem, I impose more stringent moment condition A12-i than that used in Section 1.4, A1. This allows me to apply a powerful method developed in [33] and replace $n\bar{A}_n^4(\log \bar{p})^7 \to 0$ 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 uniformly over $\mathcal{M}_{NP}$ the following assumptions hold: A4 with s and $\mathcal{S}_n$ replaced by $\bar{s}$ and $\bar{\mathcal{S}}_n$, A12, either (A2 with $\widehat{f}(X_i)$ and $f(X_i)$ replaced by $\widehat{f}(X_i, Z_i)$ and $f(X_i, Z_i)$ and $\log \bar{p} = o(n^{\kappa_1 \wedge \kappa_3})$) or (A3 and $\log \bar{p} = o(n^{\kappa_2 \wedge \kappa_3})$). The following theorem shows that the test in a multivariate nonparametric model controls asymptotic size.

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*

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

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

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

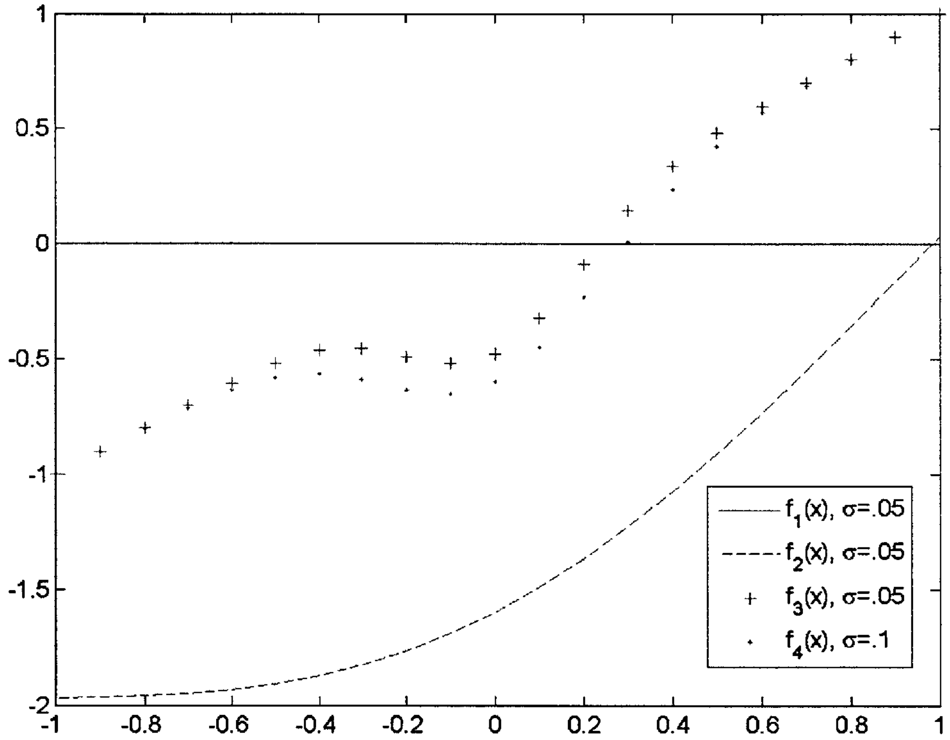
1.7 Monte Carlo Simulations

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 [50] (GHJK), [49] (GSV), and [57] (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 continuous 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 $H_n = \{h_{\max} u^l : h \geqslant h_{\min}, l = 0, 1, 2, ...\}$, $u = 0.5$, $h_{\max} = 1$, $h_{\min} = 0.4 h_{\max} (\log n/n)^{1/3}$, 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 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 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

Figure 1-2: Regression Functions Used in Simulations



precisely, I consider the test with σ_i estimated by the Rice's method (see equation (1.6)), 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 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.

Table 1.1: Results of Monte Carlo Experiments

N	C	Sample	Proportion of Rejections for								
			GSV	GHJK	HH	CS-PI	CS-OS	CS-SD	IS-PI	IS-OS	IS-SD
n	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
n	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
n	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
n	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
u	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
u	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
u	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
u	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. N and C in the heading refer to "Noise" and "Case", respectively. GSV, GHJK, and HH stand for the tests of [49], [50], and [57] 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.

1.8 Empirical Application

In this section, I review the arguments of [44] 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, [44] 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 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

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

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

[7] In the original paper, [44] 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 1.2: Incumbent Behavior versus Market Size: Monotonicity Test p-value

	Investment Type					
Method	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

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 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 [44].

1.9 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 [44]. 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.

1.10 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.

1.10.1 Plug-in Approach

1. For each $b = \overline{1,B}$ and $i = \overline{1,n}$, calculate $Y^{\star}_{i,b} = \widehat{\sigma}_i \epsilon_{i,b}$.

2. For each $b = \overline{1,B}$, calculate the value $T^{\star}_b$ of the test statistic using the sample $\{X_i, Y^{\star}_{i,b}\}_{i=1}^n$.

3. Define the plug-in critical value, $c^{PI}_{1-\alpha}$, as the $(1-\alpha)$ sample quantile of $\{T^{\star}_b\}_{b=1}^B$.

1.10.2 One-Step Approach

1. For each $b = \overline{1,B}$ and $i = \overline{1,n}$, calculate $Y^{\star}_{i,b} = \widehat{\sigma}_i \epsilon_{i,b}$.

2. Using the plug-in approach, simulate $c^{PI}_{1-\gamma}$.

3. Define $\mathcal{S}^{OS}_n$ as the set of values $s \in \mathcal{S}_n$ such that $b(s)/(\widehat{V}(s))^{1/2} > -2c^{PI}_{1-\gamma}$.

4. For each $b = \overline{1,B}$, calculate the value $T^{\star}_b$ of the test statistic using the sample $\{X_i, Y^{\star}_{i,b}\}_{i=1}^n$ and taking maximum only over $\mathcal{S}^{OS}_n$ instead of $\mathcal{S}_n$.

5. Define the one-step critical value, $c^{OS}_{1-\alpha}$, as the $(1-\alpha)$ sample quantile of $\{T^{\star}_b\}_{b=1}^B$.

1.10.3 Stepdown Approach

1. For each $b = \overline{1,B}$ and $i = \overline{1,n}$, calculate $Y^{\star}_{i,b} = \widehat{\sigma}_i \epsilon_{i,b}$.

2. Using the plug-in and one-step approaches, simulate $c^{PI}_{1-\gamma}$ and $c^{OS}_{1-\gamma}$, respectively.

3. Denote $\mathcal{S}^0_n = \mathcal{S}^{OS}_n$, $c^0 = c^{OS}_{1-\gamma}$, and set $l = 0$.

4. For given value of $l \geqslant 0$, define $\mathcal{S}^{l+1}_n$ as the set of values $s \in \mathcal{S}^l_n$ such that $b(s)/(\widehat{V}(s))^{1/2} > -c^{PI}_{1-\gamma} - c^l$.

5. For each $b = \overline{1,B}$, calculate the value $T^{\star}_b$ of the test statistic using the sample $\{X_i, Y^{\star}_{i,b}\}_{i=1}^n$ and taking the maximum only over $\mathcal{S}^{l+1}_n$ instead of $\mathcal{S}_n$.

6. Define c^{l+1}, as the $(1-\gamma)$ sample quantile of $\{T^{\star}_b\}_{b=1}^B$.

7. If $\mathcal{S}^{l+1}_n = \mathcal{S}^l_n$, 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 $\{T_b^\star\}_{b=1}^B$.

1.11 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 8 establishes existence of such a sequence under Assumptions A1, A3, A4, and A5 and Lemma 12 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 $\|g^{(r)}\|_\infty = \sup_{t\in\mathbb{R}} |g^{(r)}(t)|$.

1.12 Appendix C. Proofs for section 1.4

In this Appendix, I first prove a sequence of auxiliary lemmas (subsection 3.5). Then I present the proofs of the theorems stated in section 1.4 (subsection 1.12.2).

1.12.1 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 [111]. □

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 $\mathcal{S} \subset \mathcal{S}_n$, $c_{\eta+\delta}^{\mathcal{S},0} - c_\eta^{\mathcal{S},0} \geqslant C\delta/(\log p)^{1/2}$ for some constant $C > 0$.*

Proof. The first claim follows by combining Lemma 4 in this paper and Theorem 3 in [34]. 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 $t \in \mathbb{R}$,*

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

Proof. Recall that $c_\eta^{\mathcal{S},0}$ is the η quantile of $\max_{s\in\mathcal{S}} e(s)$, and so combining Lemma 4 and Markov inequality shows that $c_\eta^{\mathcal{S},0} \lesssim (\log p)^{1/2}/(1-\eta)$. Therefore, Lemma 32 gives

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

if $C > 0$ is sufficiently large. The lower bound follows similarly. □

Lemma 7. *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. Note that $\sum_{1\leqslant i\leqslant n}(a_i(s)\sigma_i)^2 = 1$. In addition, it follows from Assumption A1 that $\sigma_i \leqslant C$ uniformly over all $i = \overline{1,n}$. Therefore, under Assumption A5, the claim of the lemma follows by applying Corollary 2.3, case E.5 in [33]. □

Lemma 8. *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 $\eta \in (0,1)$, $\mathrm{P}(c^{\mathcal{S},0}_{\eta+\psi_n} < c^{\mathcal{S}}_\eta) = o(1)$ and $\mathrm{P}(c^{\mathcal{S}}_{\eta+\psi_n} < c_\eta^{\mathcal{S},0}) = o(1)$.*

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

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

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

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

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

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 32. 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,

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

where the last line uses Lemma 32. Therefore, on A_n, $c_\eta^{\mathcal{S}} \leqslant c_{\eta+\psi_n}^{\mathcal{S},0}$, i.e. $\mathrm{P}(c_{\eta+\psi_n}^{\mathcal{S},0} < c_\eta^{\mathcal{S}}) = o(1)$. The second claim follows similarly. □

Lemma 9. *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 $\mathrm{P}(c_{\eta+\tilde{\psi}_n}^{\mathcal{S},0} < c_\eta^{\mathcal{S},1}) = o(1)$ and $\mathrm{P}(c_{\eta+\tilde{\psi}_n}^{\mathcal{S},1} < c_\eta^{\mathcal{S},0}) = o(1)$ uniformly over $\mathcal{S} \subset \mathcal{S}_n$ and $\eta \in (0,1)$.*

Proof. I will invoke the following result recently obtained by [34].

Lemma 10. *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 $\Delta_\Sigma = \max_{1\leqslant j,k\leqslant p} |\Sigma_{jk}^1 - \Sigma_{jk}^2|$.

Proof. See Theorem 1 and following comments in [34]. □

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 Gaussian p-vectors with covariances Σ^1 and Σ^2 given by

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

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

Lemma 11. $(\log p)^2\Delta_\Sigma = o_p(1)$.

Proof. Let $u = u_n = n^{1/4}$. 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) follows from the fact that $\sum_{1 \leqslant i \leqslant n} a_i(s)^2\sigma_i^2 = 1$ by construction, and (4) is by Assumption A5. Therefore,

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

Note that $|a_i(s_1)a_i(s_2)(\tilde{\varepsilon}_i^2 - \tilde{\sigma}_i^2)| \leqslant 2A_n^2u^2$. 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 [111]) gives for some $C > 0$,

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

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 $\log p = o(1/((\log p)^4 A_n^2))$ (to verify it, note that $nA_n^4(\log p)^7 = o(1)$ implies that $n^{1/2}A_n^2(\log p)^3 = o(1)$ and $\log p/n^{1/4} = o(1)$ implies that $(\log p)^2/n^{1/2} = o(1)$; multiplying these equations gives $A_n^2(\log p)^5 = o(1)$) and $\log p = o(1/((\log p)^2 A_n^2 u^2))$, which follows from $nA_n^4(\log p)^7 = o(1)$. □

It follows from Lemma 11 that there exists a sequence $\{\tilde\psi_n\}$ of positive numbers converging to zero such that

$$(\log p)^2 \Delta_\Sigma = o_p(\tilde\psi_n^4). \tag{1.11}$$

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

$$\begin{aligned}
\|g_n^{(1)}\|_\infty &\lesssim 1/(c^{\mathcal{S},0}_{\eta+\tilde\psi_n} - c^{\mathcal{S},0}_{\eta+\tilde\psi_n/2}) \lesssim (\log p)^{1/2}/\tilde\psi_n,\\
\|g_n^{(2)}\|_\infty &\lesssim 1/(c^{\mathcal{S},0}_{\eta+\tilde\psi_n} - c^{\mathcal{S},0}_{\eta+\tilde\psi_n/2})^2 \lesssim (\log p)/\tilde\psi_n^2.
\end{aligned}$$

Applying Lemma 10 gives

$$D_n = |\mathrm{E}[g_n(\max_{s\in\mathcal{S}} Z_s^1) - g_n(\max_{s\in\mathcal{S}} Z_s^2)|\{\varepsilon_n\}]| \lesssim (\log p)\Delta_\Sigma/\tilde\psi_n^2 + (\log p)(\Delta_\Sigma)^{1/2}/\tilde\psi_n = o_p(\tilde\psi_n) \tag{1.12}$$

by equation (1.11). Note that $\max_{s\in\mathcal{S}} Z_s^1 = T^{\mathcal{S},1}$ and, using the notation of the proof of Lemma 8, $\max_{s\in\mathcal{S}} Z_s^2 = T^{\mathcal{S},0}$. Then

$$\mathrm{P}(T^{\mathcal{S},1} \leqslant c^{\mathcal{S},0}_{\eta+\tilde\psi_n}|\{\varepsilon_i\}) \geqslant_{(1)} \mathrm{E}[g_n(T^{\mathcal{S},1})|\{\varepsilon_i\}] \geqslant_{(2)} \mathrm{E}[g_n(T^{\mathcal{S},0})|\{\varepsilon_i\}] - D_n$$

$$\geqslant_{(3)} \mathrm{P}(T^{\mathcal{S},0} \leqslant c^{\mathcal{S},0}_{\eta+\tilde\psi_n/2}|\{\varepsilon_i\}) - D_n \tag{1.13}$$

$$=_{(4)} \mathrm{P}(T^{\mathcal{S},0} \leqslant c^{\mathcal{S},0}_{\eta+\tilde\psi_n/2}) - D_n \geqslant \eta + \tilde\psi_n/2 - D_n \tag{1.14}$$

where (1) and (3) are by construction of the function g_n, (2) is by equation (1.12), 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 (1.14) 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 12. *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 9 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

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

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

Note that $|T^{\mathcal{S}} - T^{\mathcal{S},1}| \leqslant p_1 + p_2$ and that by Lemmas 4 and 9, $\max_{s\in\mathcal{S}} |\sum_{1\leqslant i\leqslant n} a_i(s)\varepsilon_i\epsilon_i| = O_p((\log p)^{1/2})$. Therefore, the result follows by the argument similar to that used in the proof of Lemma 8 since $\widehat{\sigma}_i - \varepsilon_i = o_p(n^{-\kappa_1})$ by assumption A2. □

Lemma 13. *Let Assumptions A1, A4, and A5 hold. In addition, let either Assumption A2 or A3 hold. Then $\mathrm{P}(\mathcal{S}^R_n \subset \mathcal{S}^{SD}_n) \geqslant 1 - \gamma_n + o(1)$ and $\mathrm{P}(\mathcal{S}^R_n \subset \mathcal{S}^{OS}_n) \geqslant 1 - \gamma_n + o(1)$.*

Proof. Suppose that $\mathcal{S}^R_n \backslash \mathcal{S}^{SD}_n \neq \emptyset$. Then there exists the smallest integer l such that $\mathcal{S}^R_n \backslash \mathcal{S}^l_n \neq \emptyset$ and $\mathcal{S}^R_n \subset \mathcal{S}^{l-1}_n$ (if $l = 1$, let $\mathcal{S}^0_n = \mathcal{S}_n$). Therefore, $c^R_{1-\gamma_n} \leqslant c^{l-1}_{1-\gamma_n}$. It

[8] Note that Lemmas 8 and 12 provide the same results under two different methods for estimating σ_i.

follows that there exists an element s of $\mathcal{S}_n^R$ such that

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

and so

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

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 32, (7) is by Lemma 7, 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 14. *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

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

where (1) follows from the definition of $\mathcal{S}_n^R$, (2) is by Lemma 7, and (3) is by the definition of ψ_n. □

1.12.2 Proofs of Theorems

Proof of Theorem 8. Note that

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

where (1) follows from Lemma 14, (2) is by Lemma 13, (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 7, and (9) is by the definitions of ψ_n and γ_n. The first asserted claim follows.

In addition, when f is identically constant,

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

where (1) follows from the fact that $\widehat{f}(s) = 0$ whenever f is identically constant, (2) follows from $\mathcal{S}_n^P \subset \mathcal{S}_n$, (3) is by the definition of ψ_n, (4) is by Assumption A4, (5) is by Lemma 6, and (6) is from Lemma 7 and the definition of ψ_n. The second asserted claim follows. $\square$

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

$$\sum_{1 \leqslant i,j \leqslant n} (f(X_i) - f(X_j))\text{sign}(X_j - X_i)Q(X_i, X_j, s) \geqslant Cn^2 \tag{1.15}$$

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 (1.15) 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 7. 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 (1.15) replaced by

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

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),

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

for some $C > 0$, and so $f(s) \geqslant C l_n h_n^{1+\beta}(nh)^{1/2}$. 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 $\{\varepsilon_i\}$ 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 [42] gives

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

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 A8 holds, $\omega_{n,l} \leqslant Cn^{1/2}h^{\beta+3/2} = C(\log n)^{1/2}$ where C in the last expression can be made arbitrarily small by selecting sufficiently small C_0. Therefore,

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

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. □

1.13 Appendix D. Proofs for Section 1.5

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 [93]),

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

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

To prove the second claim, let $U_n = [1/(2C_3 n^{-1/3})] + 1$ where $[\cdot]$ denotes the largest integer that is smaller or equal than the quantity inside the brackets. Let $s_l = x_{n,0} < x_{n,1} < \dots < x_{n,U_n} = s_r$ 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 assumptions, there exist constants $\underline{C}$ and $\overline{C}$ such that $\underline{C} h_{n0} \leqslant p_{n,u} \leqslant \overline{C} h_{n0}$ for all $u \in \overline{1, U_n}$. 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 (see, for example, Lemma 2.2.9 in [111]) 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,

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

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 $n \geqslant N$, $|\{i = \overline{1,n} : X_i \in [x_{n,u-1}, x_{n,u}]\}| \leqslant C_6 nh_{n0}$ 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 $u_{n,B} = Bn^{1/(4+\phi)}$. 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,

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

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), and (5) follows from Assumption (iii). In addition, exponential concentration inequality for functions with bounded differences (see, for example, Theorem 12 in [20]) 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 $t = n^{-\kappa_2'}$ yields

$$\mathrm{P}(\max_{1\leqslant i\leqslant n}|\widehat{\sigma}_i^2-\sigma_i^2| > C(b_n+(nb_n)^{-1/2}+n^{-\kappa_2'})|A_{n,B}) \lesssim \exp(\log n-n^{-2\kappa_2'+\phi/(4+\phi)}b_n) = o(1)$$

for any given $B > 0$ where the last equality follows by Assumption (ii). Therefore, $\max_{1\leqslant i\leqslant n}|\widehat{\sigma}_i^2-\sigma_i^2| = O_p(b_n+(nb_n)^{-1/2}+n^{-\kappa_2'})$. Finally, since σ_i is bounded from above and away from zero uniformly over i by Assumption A1, it follows that $\max_{1\leqslant i\leqslant n}|\widehat{\sigma}_i - \sigma_i| = O_p(b_n + (nb_n)^{-1/2} + n^{-\kappa_2'})$, 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

$$2C_6(1-\bar{C}_2)^{k+1}\max_{t\in[-1,-\bar{C}_2]}K(t) \leqslant C_5\bar{C}_2^k\bar{C}_1\min_{t\in[0,\bar{C}_1]}K(t) \tag{1.16}$$

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

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

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

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 $X_i \leqslant x - \bar{C}_2 h$, (2) is by Assumption A8, (3) is by equation (1.16), and (4) is because $\min_{t\in[0,\bar{C}_1]} K(t) > 0$. Therefore, denoting $M_n(x, h) = \{i = \overline{1,n} : X_i \in [x-(1+\bar{C}_2)h/2, x-\bar{C}_2 h]\}$, the fact that $V(s) \geqslant C(nh)^3h^{2k}$

follows from Assumptions A1 and A8 and the following calculations:

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

where $C > 0$ does not depend on (x, h). Therefore, claim (a) follows since

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

Further, under Assumption A3,

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

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

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

Consider $I_1(s)$. As in the proof of Lemma 11, let $u = u_n = n^{1/4}$. Let $\tilde{\varepsilon}_i = \varepsilon_i 1\{|\varepsilon_i| \leqslant u\}$ and $\tilde{\sigma}_i^2 = \mathrm{E}[\tilde{\varepsilon}_i^2]$. It follows from Assumption A1 that $\mathrm{P}(\max_{1\leqslant i\leqslant n} |\tilde{\varepsilon}_i - \varepsilon_i| = 0) \to 1$, 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

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

Applying Bernstein inequality and using the union bound yields

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

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,

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

and so $I_2(s)/V(s) = o_p(n^{-\kappa_1})$ uniformly over $s \in \mathcal{S}_n$. Combining presented results gives the asserted claim (c). □

1.14 Appendix E. Proofs for Section 1.6

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

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

The result follows by the argument similar to that used in the proof of Theorem 8. □

Proof of Theorem 7. The proof relies on the same notation as introduced in Section 1.11 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, respectively.

With this new notation, Lemmas 4, 32, 6, 8, 12, 13, and 14 follow without any further changes. Further, Lemma 7 now follows by applying Corollary 2.3, case E.4 in [33]. The proof of Lemma 9 is the same as before (in particular, inner lemma 10 does not require any changes), with the exception that now in the proof of inner lemma 11, I set $u = u_n = C\log(\bar{p}n)$ for sufficiently large C and $\phi = 1$; then conclusion of Lemma 11, which requires $(\log \bar{p})^2/u^{2+\phi} = o(1)$, $\log \bar{p} = o(1/((\log \bar{p})^4 A_n^2))$ and $\log \bar{p} = o(1/((\log \bar{p})^2 A_n^2 u^2))$, follows from imposed condition that $A_n^2(\log(\bar{p}n))^7 = o(1)$.

Now, the first claim of Theorem 7 follows from an argument similar to that used in the proof of Theorem 8 by noting that under $\mathcal{H}_0$, $\max_{\bar{s} \in \bar{\mathcal{S}}_n} \widehat{f}(\bar{s}) \leqslant o_p(1/\sqrt{\log \bar{p}})$[9],

[9] Specifically, the only required change in the proof of the first claim of Theorem 8 is that starting

which holds by conditions (iii) and (iv) of Assumption A12.. The second claim of the theorem again follows from an argument similar to that used in the proof of Theorem 8 by noting that when f is identically constant, $\max_{\bar{s}\in\bar{S}_n}|\widehat{f}(\bar{s})| \leqslant o_p(1/\sqrt{\log \bar{p}})$, which holds by conditions (iii) and (iv) of Assumption A12.

□

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 (1.2) with $k = 1$. The simulation design is the same as in Section 1.7. 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.

from inequality (3) α should be replaced by $\alpha - o(1)$.

Table 1.3: Results of Monte Carlo Experiments

N	C	Sample	Proportion of Rejections for								
			GSV	GHJK	HH	CS-PI	CS-OS	CS-SD	IS-PI	IS-OS	IS-SD
n	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
n	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
n	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
n	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
u	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
u	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
u	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
u	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. N and C in the heading refer to "Noise" and "Case", respectively. GSV, GHJK, and HH stand for the tests of [49], [50], and [57] 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.

Chapter 2

Adaptive Test of Conditional Moment Inequalities

2.1 Introduction

Conditional moment inequalities (CMI) are important both in economics and in econometrics. In economics, they arise naturally in many models that include behavioral choice, see [91] for a survey. In econometrics, they appear in estimation problems with interval data and problems with censoring; see, for example, [80]. In addition, CMI offer a convenient way to study treatment effects in randomized experiments as described in [74]. In the next section, I provide three detailed examples of models with CMI.

To describe CMI model, let $m : \mathbb{R}^d \times \mathbb{R}^k \times \Theta \to \mathbb{R}^p$ be a vector-valued known function. Let (X, W) be a pair of $\mathbb{R}^d$ and $\mathbb{R}^k$-valued random vectors, and $\theta \in \Theta$ a parameter. Then CMI are given by the following equation:

$$E[m(X, W, \theta)|X] \leqslant 0 \, a.s. \tag{2.1}$$

Note that (2.1) also covers conditional moment equalities (CME) because CME can be represented as pairs of CMI. In this paper, I am interested in testing the null hypothesis, H_0, that $\theta = \theta_0$ against the alternative, H_a, that $\theta \neq \theta_0$ based on a random sample $\{(X_i, W_i)\}_{i=1}^n$ from the distribution of (X, W).

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

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

denote the identified set. CMI are said to identify θ if and only if Θ_I is a singleton. Otherwise, CMI do not identify the parameter θ. For example, non-identification may happen when the CMI arise from a game-theoretic model with multiple equilibria. Moreover, the parameter may be weakly identified, which means that Θ_I is a singleton but information on Θ_I contained in the data is limited even in large samples. My approach leads to a robust test with the correct asymptotic size no matter whether the parameter θ is identified, weakly identified, or not identified. Here "robust" means that the test controls asymptotic size uniformly over a large class of models.

Testing CMI has been a popular research topic in econometrics recently. As a result, two approaches to robust CMI testing have been developed. One approach ([5]), is based on converting CMI into an infinite number of unconditional moment inequalities using nonnegative weighting functions. The other approach ([36]), is based on estimating moment functions nonparametrically.

To motivate the test developed in this paper, consider two (highly stylized) examples of CMI models. Even though these models are very simple, they convey the main ideas. In the first model, m is multiplicatively separable in θ, i.e.

$$m(X, W, \theta) = \theta \tilde{m}(X, W)$$

for some $\tilde{m} : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ and $\theta \in \mathbb{R}$ with $E[\tilde{m}(X, W)|X] > 0$ a.s. In the second model, m is additively separable in θ, i.e.

$$m(X, W, \theta) = \tilde{m}(X, W) + \theta.$$

The identified sets, Θ_I, in these models are

$$\{\theta \in \mathbb{R} : \theta \leqslant 0\} \text{ and } \{\theta \in \mathbb{R} : \theta \leqslant -\operatorname*{ess\,sup}_X E[\tilde{m}(X, W)|X]\}$$

correspondingly [1]. [5] developed a test that has nontrivial power against alternatives of the form $\theta_0 = \theta_{0,n} = C/\sqrt{n}$ for any $C > 0$ in the first model, and so their test has extremely high power in this model. It follows from [10], however, that (in comparison with the test of [36]), the test of [5] often has low power in the second model[2]. The purpose of this paper is to construct a test that has high power in a large class of CMI models including models like that in the second example, and at the same time, that has nearly the same power as that of [5] in models like that described in the first example. The key difference of my approach is that my test statistic is based on the *studentized* estimates of moments whereas theirs is not. More precisely, [5] consider studentized statistics but modify the variance term so that asymptotic power properties of their test are similar to those of the test with no studentization.

The test of [36] also has high power in a large class of CMI models but implementing their test requires knowledge of certain smoothness properties of moment functions whereas the test developed in this paper does not require this information. Moreover, my test automatically adapts to these smoothness properties selecting the most appropriate weighting function. For this reason, I call my test adaptive. This feature of the test is important because smoothness properties of moment functions are rarely known in practice. On the other hand, an advantage of the test of [36] is that it becomes very efficient if moment functions are sufficiently smooth (for example, if moment functions are at least twice continuously differentiable).[3,4]

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

[2] [5] developed tests based on both Cramer-von Mises and Kolmogorov-Smirnov test statistics. In this paper, I refer to their test with the Kolmogorov-Smirnov test statistic. Most statements are also applicable for Cramer-von Mises test statistic as well, however.

[3] Efficiency of the test of [36] is achieved by using higher order kernel or series methods for estimating moment functions; both the test of [5] and the test developed in this paper work with positive kernels, which exclude higher order kernels, and do not have this feature of the test of [36].

[4] In the statistics literature, there recently have been developed techniques for adaptively selecting the appropriate smoothing parameter for tests like that in [36]. An example is Lepski's method combined with the sample splitting where a part of the sample is used to select the smoothing

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

I develop a bootstrap method to simulate a critical value for the test. The method is based on the observation that the distribution of the test statistic in large samples depends on the distribution of the noise $\{m(X_i, W_i, \theta_0) - E[m(X_i, W_i, \theta_0)|X_i]\}_{i=1}^n$ only via second moments of the noise. For reasons similar to those discussed in [35] and [6], the distribution of the test statistic in large samples depends heavily on the extent to which the CMI are binding. Moreover, the parameters that measure to what extent the CMI are binding cannot be estimated consistently. Therefore, I develop a new approach to deal with this problem, which I refer to as the refined moment selection (RMS) procedure. The approach is based on a pretest which is used to decide what counterparts of the test statistic should be used in simulating the critical value for the test. Unlike [5], I use a model-specific, data-driven, critical value for the pretest, which is taken to be a large quantile of the appropriate distribution, whereas they use a deterministic threshold with no reference to the model. I also provide a plug-in critical value for the test. My proof of the bootstrap validity is nonstandard because it uses only finite sample arguments. My proof is also different from those used in [5] and [36].

None of the tests in the literature including mine have power against alternatives

parameter according to the Lepski's algorithm and the other part is used for testing; see, for example, [51]. Deriving formal results on how the test of [36] works in combination with these adaptive smoothing parameter selection techniques would be an important direction for future research.

in the set Θ_I. Therefore, I consider the alternatives of the form

$$P(E[m_j(X,W,\theta_0)|X] > 0) > 0 \text{ for some } j = 1,...,p. \tag{2.2}$$

To show that my test has good power properties in a large class of CMI models, I derive its power against alternatives of the form (2.2) assuming that $E[m(X,W,\theta_0)|X]$ is some vector of unrestricted nonparametric functions. In other words, I consider nonparametric classes of alternatives. Once $m(X,W,\theta)$ is specified, it is straightforward to translate my results to the parametric setting. The test developed in this paper is consistent against any fixed alternative outside of the set Θ_I. I also show that my method allows for nontrivial testing against $(n/\log n)^{-1/2}$-local one-dimensional alternatives.[5] Finally, I prove that the test is minimax rate optimal against certain classes of smooth alternatives consisting of moment functions $E[m(X,W,\theta_0)|X]$ that are sufficiently flat at the points of maxima. Minimax rate optimality means that the test is uniformly consistent against alternatives in the mentioned class whose distance from the set of models satisfying (2.1) converges to zero at the fastest possible rate. The requirement that functions should be sufficiently flat cannot be dropped because of the restrictions on kernels used for estimating moment functions.

One of the advantages of the proof technique used in this paper is that it gives an explicit error bound on the bootstrap approximation error of the distribution of the test statistic. In particular, it allows me to show that for the test developed in this paper, the probability of rejecting the null under the null can exceed nominal level only by a *polynomially* (in n) small term. In contrast, all other papers on CMI only show that this probability is *asymptotically* not larger than the nominal level. I believe that this contribution is important in light of the fact that asymptotic approximations of suprema of processes typically provide only *logarithmically* (in n) small approximation error (see, for example, [56]).

The literature concerned with unconditional and conditional moment inequalities

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

is expanding quickly. Published papers on unconditional moment inequalities include [35], [101], [104], [3], [4], [6], [23], [26], [91], and [102]. There is also a large literature on partial identification which is closely related to that on moment inequalities. Methods specific for conditional moment inequalities were developed in [65], [66], [36], [5], [74], [10], [11], and [12]. The case of CMI that point identify θ is treated in [65]. The test of [66] is closely related to that of [5]. [74] developed a test based on the minimum distance statistic in the one-sided L_p-norm and kernel estimates of moment functions. The advantage of their approach comes from simplicity of their critical value for the test, which is an appropriate quantile of the standard Gaussian distribution. Their test is not adaptive, however, since only one bandwidth value is used. [10] developed a new method for computing the critical value for the test statistic of [5] that leads to a more powerful test than theirs but the resulting test is not robust. In particular, his method cannot be used in CMI models like that described in the first example above. [11], which was written independently and at the same time as this paper, considered a test statistic similar to that used in this paper and derived a critical value such that the whole identified set is contained in the confidence region with probability approaching one. In other words, he focused on estimation rather than inference. In addition, the critical value in that paper is infeasible since it has the form $a_n\sqrt{\log n/n}$ where a_n is some sequence of unknown positive numbers that is bounded away from zero. After this paper had been made publicly available, [12] constructed a test based on a test statistic that is closely related to that used in this paper with the critical value derived from the limit distribution.[6]

Finally, an important related paper in the statistical literature is [42]. They consider testing qualitative hypotheses in the ideal Gaussian white noise model where a researcher observes a stochastic process that can be represented as a sum of the mean function and a Brownian motion. In particular, they developed a test of the

[6] It should be noted, however, that [12] imposed rather strong assumptions. Specifically, they assumed existence of finite moment generating function of the noise. In contrast, I only assume 4 finite moments, which is much more plausible in applications. In addition, they assumed that moment functions are not too closely related with each other. Finally, they only provided a plug-in critical value, and it is not obvious how to extend their methods to derive a moment selection procedure.

hypothesis that the mean function is (weakly) negative almost everywhere. Though their test statistic is somewhat related to that used in this paper, the technical details of their analysis are quite different.

The rest of the paper is organized as follows. The next section discusses some examples of CMI models. Section 2.3 formally introduces the test. The main results of the paper are presented in Section 2.4. Extensions to the cases of infinitely many CMI and local CMI are provided in Section 2.5. A Monte Carlo simulation study is described in Section 2.6. There I provide an example of an alternative with a well-behaved moment function such that the test developed in this paper rejects the null hypothesis with probability higher than 80% while the rejection probability of all previous tests does not exceed 20%. Brief conclusions are drawn in Section 2.7. Finally, all proofs are contained in the Appendix.

2.2 Examples

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

Incomplete Models of English Auctions. My first example follows [55] treatment of English auctions under weak conditions. The popular model of English auctions suggested by [83] assumes that each bidder is holding down the button while the price for the object is going up continuously until she wants to drop out. The price at the moment of dropping out is her bid. It is well-known that the dominant strategy in this model is to make a bid equal to her valuation of the object. In practice, participants usually call out bids, however. Hence, the price rises in jumps, and the bid may not be equal to person's valuation of the object. In this situation, the relation between bids and valuations of the object depends crucially on the modeling assumptions. [55] derived certain bounds on the distribution function of valuations based on minimal assumptions of rationality.

Suppose that we have an auction with m bidders whose valuations of the object are drawn independently from the distrubution $F(\cdot, X)$ where X denotes observable characterics of the object. Let $b_1, ..., b_m$ denote highest bids of each bidder. Let

$b_{1:m} \leqslant ... \leqslant b_{m:m}$ denote the ordered sequence of bids $b_1, ..., b_m$. Assuming that bids do not exceed bidders' valuations, [55] derived the following upper bound on $F(\cdot, X)$:

$$E[\phi^{-1}(F(v, X)) - I\{b_{i:m} \leqslant v\}|X] \leqslant 0\, a.s. \tag{2.3}$$

for all $v \in \mathbb{R}$ and $i = 1, ..., m$ where $\phi(\cdot)$ is a certain (known) increasing function, see equation (3) in [55]. A similar lower bound follows from the assumption that bidders do not allow opponents to win at a price they would like to beat. Parameterizing the function $F(\cdot, \cdot)$ and considering (2.3) for a finite set $\mathcal{V} = \{v_1, ..., v_p\}$ of values of v gives inequalities of the form (2.1).

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

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

where $E[\varepsilon|X, V] = 0$ a.s. and V is a scalar random variable. Suppose that we observe X and Y but do not observe V. Instead, we observe V_0 and V_1, called brackets, such that $V \in [V_0, V_1]$ a.s. In empirical analysis, brackets may arise because a respondent refuses to provide information on V but provides an interval to which V belongs. Following [80], assume that $f(X, V)$ is weakly increasing in V and $E[Y|X, V] = E[Y|X, V, V_0, V_1]$. Then it is easy to see that

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

and

$$E[I\{V_0 \geqslant v\}(Y - f(X, v))|X, V_0, V_1] \geqslant 0 \tag{2.5}$$

for all $v \in \mathbb{R}$. Again, parameterizing the function $f(\cdot, \cdot)$ and selecting a finite set $\mathcal{V} = \{v_1, ..., v_p\}$ gives inequalities of the form (2.1).

Treatment Effects. Suppose that we have a randomized experiment where one group of people gets a new treatment while the control group gets a placebo. Let $D = 1$ if the person gets the treatment and 0 otherwise. Let p denote the probability that

$D = 1$. Let X denote person's observable characteristics and Y denote the realized outcome. Finally, let Y_0 and Y_1 denote the counterfactual outcomes had the person received a placebo or the new medicine respectively. Then $Y = DY_1 + (1 - D)Y_0$. The question of interest is whether the new medicine has a positive expected impact uniformly over all possible charactersics X. In other words, the null hypothesis, H_0, is that

$$E[Y_1 - Y_0|X] \geqslant 0\, a.s. \tag{2.6}$$

Since in randomized experiments D is independent of X, [74] showed that

$$E[Y_1 - Y_0|X] = E[DY/p - (1 - D)Y/(1 - p)|X]. \tag{2.7}$$

Combining (2.6) and (2.7) gives CMI of the form (2.1).

2.3 Test

In this section, I present the test statistic and give two bootstrap methods to simulate critical values. The analysis in this paper is conducted conditional on the set of values $\{X_i\}_{i=1}^{\infty}$, so all probabilistic statements excluding those in Lemmas 17 and 18 in the Appendix should be understood conditional on $\{X_i\}_{i=1}^{\infty}$ for almost all sequences $\{X_i\}_{i=1}^{\infty}$. Lemmas 17 and 18 provide certain conditions that ensure that the assumptions used in this paper hold for almost all sequences $\{X_i\}_{i=1}^{\infty}$.

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

$$f(X) \leqslant 0\, a.s.$$

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

Section 2.3.1 defines the test statistic assuming that $\Sigma_i = E[\varepsilon_i \varepsilon_i^T|X_i]$ is known for each $i = 1, ..., n$. Section 2.3.2 gives two bootstrap methods to simulate critical values. The first one is based on plug-in asymptotics, while the second one uses the refined moment selection (RMS) procedure. Section 2.3.2 also provides some intuition of why these procedures lead to the correct asymptotic size of the test. When Σ_i is unknown, it should be estimated from the data. Section 2.3.3 shows how to construct an appropriate estimator $\widehat{\Sigma}_i$ of Σ_i. The feasible version of the test will be based on substituting $\widehat{\Sigma}_i$ for Σ_i both in the test statistic and in the critical value.

2.3.1 The Test Statistic

The test statistic in this paper is based on a kernel estimator of the vector-valued function f. Let $K : \mathbb{R}^d \to \mathbb{R}_+$ be some kernel. For bandwidth value $h \in \mathbb{R}_+$, let $K_h(x) = K(x/h)/h^d$. For each pair of observations $i, j = 1, ..., n$, denote the weight function

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

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

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

where $Y_{j,m}$ denotes m-th component of Y_j[7]. Conditional on $\{X_i\}_{i=1}^n$, the variance of the kernel estimator $\widehat{f}_{(i,m,h)}$ is

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

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

Next, consider a finite set of bandwidth values $H = \{h = h_{\max} a^k : h \geqslant h_{\min}, k = 0, 1, 2, ...\}$ for some $h_{\max} > h_{\min}$ and $a \in (0, 1)$. For simplicity, I assume that $h_{\min} = h_{\max} a^k$ for some $k \in \mathbb{N}$ so that $h_{\min}$ is included in H. I assume that as the sample size n increases, $h_{\min}$ converges to zero while $h_{\max}$ is bounded away from zero. For practical purposes, I recommend setting $K(x) = 0.75(1 - \|x\|^2)$ for $\|x\| \leqslant 1$ and 0 otherwise,

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

$h_{\max} = \max_{i,j=1,\ldots,n} \|X_i - X_j\|/2$, $h_{\min} = 0.2h_{\max}(\log n/n)^{1/(3d)}$, and $a = 0.5$.[8] This choice of parameters is consistent with the theory presented in the paper and also worked well in my simulations. Note that $h_{\min}$ is chosen so that the kernel estimator uses on average roughly 15 data points when $n = 250$.

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

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

Thus, the test statistic is based on the studentized kernel estimates of the function f at points $\{X_i : i = 1, ..., n\}$.[9]

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

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

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

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

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

instead of T, where $\widehat{\Sigma}_j$ is some estimator of Σ_j. Some possible estimators are discussed in Section 2.3.3.

2.3.2 Critical Values

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

The first method relies on three observations. First, one can approximate $\widehat{T}$ by T. Second, it is easy to see that, for a fixed distribution of disturbances $\{\varepsilon_i\}_{i=1}^n$, the maximum of $(1-\alpha)$ quantile of the test statistic T over all possible functions f satisfying $f \leqslant 0_p$ a.s. corresponds to $f = 0_p$. Third, Lemma 25 in the Appendix shows that the distribution of T is asymptotically independent of the distrubution of disturbances $\{\varepsilon_i : i = 1, ..., n\}$ apart from their second moments $\{\Sigma_i : i = 1, ..., n\}$. These observations suggest that one can simulate $c_{1-\alpha}$ (denoted by $c_{1-\alpha}^{PIA}$) by the following procedure:

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

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

3. Repeat steps 1 and 2 independently B times for some large B to obtain $\{T_b^{PIA} : b = 1, ..., B\}$.

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

The second method is based on the refined moment selection (RMS) procedure. It gives a more powerful test and still controls the required size. The method relies on the observation that $|\widehat{T}| = O_p(\sqrt{\log n})$ if $f = 0_p$ (see Lemmas 22, 23, and 25 in the Appendix) while $\widehat{f}_{i,m,h}/\widehat{V}_{(i,m,h)} \to -\infty$ at a polynomial rate if $f_m(X) < 0$ for X satisfying $\|X - X_i\| < h$. Such terms will have asymptotically negligible effect on the distribution of $\widehat{T}$, so we can ignore corresponding terms in the simulated statistic. Therefore, one can simulate $c_{1-\alpha}$ (denoted by $c_{1-\alpha}^{RMS}$) as follows. First, let $\gamma < \alpha/2$ be some small positive number (truncation parameter). Second, use the plug-in bootstrap to find $c_{1-\gamma}^{PIA}$. Denote

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

Third, run the following procedure:

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

In the next section, it will be assumed that $\gamma = \gamma_n \to 0$ as $n \to \infty$. So, I recommend setting γ as a small fraction of α, for example $\gamma = 0.01$ for $\alpha = 0.05$. Alternatively, one can set $\gamma = 0.1/\log(n)$, similar to [36]. [10]

2.3.3 Estimating Σ_i

Let me now explain how one can estimate Σ_i. The literature on estimating Σ_i is huge. Among other papers, it includes [97], [88], [59], and [46]. For scalar-valued Y_i, available estimators are described in [63]. All those estimators can be immediately generalized to vector-valued Y_i's. For concreteness, I describe one estimator here. Choose a bandwidth value $b_n > 0$. For $i = 1, ..., n$, let $J(i) = \{j = 1, ..., n : \|X_j - X_i\| \leqslant b_n\}$. If $J(i)$ has an odd number of elements, drop one arbitrarily selected observation. Partition $J(i)$ into pairs using a map $k : J(i) \to J(i)$ satisfying $k(j) \neq j$ and

[10] Note also that if γ is comparable with α, one can do a finite sample adjustment of the critical value by taking $(1 - \alpha + 2\gamma)$ quantile of $\{T_b^{RMS}\}_{b=1}^{B}$ at step 4 of the procedure above. Also, the theory in the next section requires that $\gamma \leqslant Cn^{-c}$ for some constants c and C. Nevertheless, in my Monte Carlo simulations I use the rule $\gamma = 0.1/\log(n)$ to make meaningful comparisons with [36].

$k(k(j)) = j$ for all $j \in J(i)$. Let $|J(i)|$ denote the number of elements in $J(i)$. Then Σ_i can be estimated by

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

Lemma 15 in the Appendix gives certain conditions that ensure that this estimator will be uniformly consistent for Σ_i over $i = 1, ..., n$ with a polynomial rate, i.e.

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

for some $\kappa > 0$ where $\|\cdot\|$ denotes the spectral norm on the space of $p \times p$-dimensional symmetric matrices corresponding to the Eucledian norm on $\mathbb{R}^p$. To choose the bandwidth value b_n in practice, one can use some version of cross validation. An advantage of this estimator is that it is fully adaptive with respect to the smoothness properties of f. Note that the estimator $\widehat{\Sigma}_i$ is based on the sample-splitting method. To improve efficiency of the estimator, one can avoid sample-splitting by using concentration inequalities as in [38]. For brevity, however, I do not consider that option here.

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

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

so that

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

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

2.4 Main Results

This section presents main results. Section 2.4.1 gives regularity conditions. Section 2.4.2 describes size properties of the test. Section 2.4.3 explains the behavior of the test under a fixed alternative. Section 2.4.4 derives the rate of consistency of the test against local one-dimensional alternatives mentioned in the Introduction. Section 2.4.5 shows the rate of uniform consistency against certain classes of smooth alternatives. Section 2.4.6 presents the minimax rate-optimality result.

2.4.1 Assumptions

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

A13. *(i) Design points* $\{X_i\}_{i=1}^n$ *are nonstochastic. (ii)* $c_1 nh^d \leqslant M_h(X_i) \leqslant C_1 nh^d$ *for all* $i = 1, ..., n$ *and* $h \in H = H_n$.

The design points are nonstochastic because the analysis is conducted conditional on X_i's. In addition, A13 states that the number of design points in certain neighborhoods of each design point is proportional to the volume of the neighborhood with the coefficient of proportionality bounded from above and away from zero. It is stated in [63] that A13 holds in an iid setting with probability approaching one as the sample size increases if the distribution of X_i is absolutely continuous with respect to Lebegue measure, has bounded support, and has density bounded away from zero on the support. This statement is not precise unless one makes some extra assumptions. Lemma 17 in the Appendix gives a counter-example. Instead, Lemma 18 shows that A13 holds for large n a.s. if, in addition, it is assumed that the density of X_i is bounded from above, and that the support of X_i is a convex set. Necessity of the density boundedness is obvious. Convexity of the support is not necessary for A13 but it strikes a good balance between generality and simplicity. In general, one must deal with some smoothness properties of the boundary of the support. Note that the statement "for large n a.s." is stronger than "with probability approaching one".

A14. *(i) Disturbances $\{\varepsilon_i\}_{i=1}^n$ are independent $\mathbb{R}^p$-valued random vectors with $E[\varepsilon_i] = 0$ for all $i = 1, ..., n$. (ii) $E[\max_{m=1,...,p} |\varepsilon_{i,m}|^4] \leqslant C_2$ for all $i = 1, ..., n$. (iii) $\Sigma_{i,mm} \geqslant c_2$ for all $i = 1, ..., n$ and $m = 1, ..., p$.*

Finite fourth moment of disturbances is used to show that the distribution of the test statistic $\widehat{T}$ in large samples does not depend on the form of the distribution of ε_i's. I assume that the variance of each component of disturbances is bounded away from zero for simplicity of the presentation. Since I use studentized kernel estimates, without this assumption, it would be necessary to truncate the variance of the kernel estimators from below with truncation level slowly converging to zero. That would complicate the derivation of the main results without changing the main ideas.

Let $\tau, L > 0$ be arbitrary positive numbers. In addition let $\varsigma \in \{1, ..., [\tau]\}$. Here $[\tau]$ denotes the largest integer strictly smaller than τ. Before stating A15, let me give formal definitions of Holder smoothness classes $\mathcal{F}(\tau, L)$ and their subsets $\mathcal{F}_\varsigma(\tau, L)$. For d-tuple of nonnegative integers $\alpha = (\alpha_1, ..., \alpha_d)$ with $|\alpha| = \alpha_1 + ... + \alpha_d$, function $g : \mathbb{R}^d \to \mathbb{R}$, and $x = (x_1, ..., x_d) \in \mathbb{R}^d$, denote

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

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

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

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

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

Let $S^{d-1} = \{l \in \mathbb{R}^d : \|l\| = 1\}$ denote the space of directions in $\mathbb{R}^d$. For any

$g \in \mathcal{F}(\tau, L)$, $x = (x_1, ..., x_d) \in \mathbb{R}^d$, and $l \in S^{d-1}$, let $g^{(k,l)}(x)$ denote k-th derivative of function g in direction l at point x whenever it exists[11]. For $\varsigma = 1, ..., [\tau]$, let $\mathcal{F}_\varsigma(\tau, L)$ denote the class of all elements of $\mathcal{F}(\tau, L)$ such that for any $g \in \mathcal{F}_\varsigma(\tau, L)$ and $l \in S^{d-1}$, $g^{(k,l)}(x) = 0$ for all $k = 1, ..., \varsigma$ whenever $g^{(1,l)}(x) = 0$, and there exist $x = (x_1, ..., x_d) \in \mathbb{R}^d$ and $l \in S^{d-1}$ such that $g^{(\varsigma+1,l)}(x) \neq 0$ and $g^{(1,l)}(x) = 0$. If $\tau \leqslant 1$, I set $\varsigma = 0$ and $\mathcal{F}_\varsigma(\tau, L) = \mathcal{F}(\tau, L)$.

A15. *Components f_m's of the regression function f satisfy $f_m \in \mathcal{F}_\varsigma(\tau, L)$ for all $m = 1, ..., p$.*

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

A16. *(i) The set of bandwidth values has the following form: $H = H_n = \{h = h_{\max} a^k : h \geqslant h_{\min}, k = 0, 1, 2, ...\}$ where $a \in (0, 1)$, $h_{\max} = \max_{i,j=1,...,n} \|X_i - X_j\|/2$ and $h_{\min} = C_3 (\log n/n)^{1/(3d)}$. (ii) $S = S_n = \{(i, m, h) : i = 1, ..., n,\ m = 1, ..., p,\ h \in H_n\}$.*

According to this assumption, the maximal bandwidth value, $h_{\max}$, is chosen to match the radius of the support of design points. It is intended to detect deviations from the null hypothesis in the form of flat alternatives. The minimal bandwidth value, $h_{\min}$, converges to zero as the sample size increases at an appropriate rate. The minimal bandwidth value is intended to detect alternatives with narrow peaks. A16(ii) is a key condition used to establish an invariance principle that shows that the distribution of $\widehat{T}$ asymptotically depends on the distribution of disturbances ε_i's only through their covariances Σ_i's.

A17. *(i) The kernel K is positive and supported on $\{x \in \mathbb{R}^d : \|x\| \leqslant 1\}$. (ii) $K(x) \leqslant 1$ for all $x \in \mathbb{R}^d$ and $K(x) \geqslant c_3$ for all $\|x\| \leqslant 1/2$.*

I assume that the kernel function is positive on its support. Many kernels satisfy this assumption. For example, one can use rectangular, triangular, parabolic, or biweight kernels. See [110] for the definitions. On the other hand, the requirement that the

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

kernel is positive on its support excludes higher-order kernels, which are necessary to achieve the minimax optimal testing rate over large classes of smooth alternatives. I require positive kernels because of their negativity-invariance property, which means that any kernel smoother with a positive kernel maps the space of negative functions into itself. This property is essential for obtaining a test with the correct asymptotic size when smoothness properties of moment functions are unknown. With higher-order kernels, one has to assume undersmoothing so that the bias of the estimator is asymptotically negligible in comparison with its standard deviation. Otherwise, large values of $\widehat{T}$ might be caused by large values of the bias term relative to the standard deviation of the estimator even though all components of $f(X)$ are negative. However, for undersmoothing, one has to know the smoothness properties of $f(X)$. In contrast, with positive kernels, the set of bandwidth values can be chosen without reference to these smoothness properties. In particular, the largest bandwidth value can be chosen to be bounded away from zero. Nevertheless, the test developed in this paper will be rate optimal in the minimax sense against classes $\mathcal{F}_{[\tau]}(\tau, L)$ when $\tau > d$.

A18. *Estimators $\widehat{\Sigma}_i$ of Σ_i satisfy* $\mathrm{P}(\max_{i=1,...,n} \|\widehat{\Sigma}_i - \Sigma_i\| > C_4 n^{-c_4}) \leqslant C_4 n^{-c_4}$ *where $\|\cdot\|$ denotes the spectral norm on the space of $p \times p$-dimensional symmetric matrices corresponding to the Euclidean norm on $\mathbb{R}^p$.*

A18 is satisfied for $\widehat{\Sigma}_i$ described in Section 2.3.3. In practice, due to the curse of dimensionality, it might be useful to use some parametric or semi-parametric estimators of Σ_i's instead of the estimator described in Section 2.3.3. For example, if we assume that $\Sigma_i = \Sigma_j$ for all $i, j = 1, ..., n$, then the estimator of [97] (or its multivariate generalization) is $1/\sqrt{n}$-consistent in L_1-norm. In this case, A18 will be satisfied with $\kappa = 1/4 - \phi$ for arbitrarily small $\phi > 0$.

A19. *The truncation parameter γ satisfies $\gamma = \gamma_n \leqslant C_5 n^{-c_5}$.*

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

A13-A15 concern the data-generating process while A16-A19 deal with the test. Taken all together, they define the model.[12] The asymptotic results in this paper will

[12] The model in this paper is understood as infinite sequences of nonstochastic design points

be shown to hold uniformly over all models satisfying A13-A19. For that purpose, the following notation will be useful. Let $\mathcal{G}$ denote the set of all models satisfying A13-A19 for all n, and let $w \in \mathcal{G}$ denote a generic model in $\mathcal{G}$. In addition, let $E_w[\cdot]$ denote the expectation calculated assuming the model w. Finally, let $f(w)$ denote the regression function f corresponding to the model w.

2.4.2 Size Properties of the Test

Analysis of size properties of the test is complicated because it is unknown whether the test statistic has a limiting distribution. Instead, I use a finite sample method developed in [33]. For each sample size n, this method gives an upper error bound on the uniform distance between the cdf of the test statistic and the cdf the test statistic would have in the model with Gaussian noise $\{\varepsilon_i\}_{i=1}^n$.

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

Theorem 8. *Let $P = PIA$ or RMS. Then for some constants c and C depending only on c_j and C_j for $j = 1, ..., 5$,*

$$\sup_{w \in \mathcal{G}_0} P_w\left(\widehat{T} > c_{1-\alpha}^P\right) \leqslant \alpha + Cn^{-c} \text{ for all } n. \tag{2.8}$$

In addition,

$$\inf_{w \in \mathcal{G}_{00}} P_w\left(\widehat{T} > c_{1-\alpha}^P\right) \geqslant \alpha - Cn^{-c} \text{ for all } n. \tag{2.9}$$

Comment 6. *(i) Proofs of all results are presented in the Appendix.*

(ii) An advantage of this theorem is that it shows that the probability of rejecting the null under the null can exceed the nominal level of the test only by a polynomially

$\{X_i\}_{i=1}^\infty$ and random disturbances $\{\varepsilon_i\}_{i=1}^\infty$, a vector-valued regression function f, sequence of estimators $\{\widehat{\Sigma}_i\}_{i=1}^n$ for each n, a kernel K, a sequence of sets of bandwidth values $\{H_n\}_{n=1}^\infty$, a sequence of sets $\{S_n\}_{n=1}^\infty$, and a sequence of truncation parameters $\{\gamma_n\}_{n=1}^\infty$.

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

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

(iv) Combining (2.8) and (2.9) shows that uniformly over $w \in \mathcal{G}_{00}$, $|P_w(\widehat{T} > c^P_{1-\alpha}) - \alpha| \leqslant Cn^{-c}$.

2.4.3 Consistency Against a Fixed Alternative

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

$$\rho(w, H_0) = \sup_{i=1,...,\infty;\, m=1,...,p} [f_m(X_i)]_+ \tag{2.10}$$

For any alternative outside of the set Θ_I, $\rho(w, H_0) > 0$. The following theorem shows that the test is consistent against any fixed alternative $w \in \mathcal{G}$ with $\rho(w, H_0) > 0$. Moreover, the theorem shows that the test is consistent uniformly against alternatives whose distance from the null hypothesis is bounded away from zero. For $\rho > 0$, let $\mathcal{G}_\rho$ denote the subset of all elements w of $\mathcal{G}$ such that $\rho(w, H_0) \geqslant \rho$.

Theorem 9. *Let $P = PIA$ or RMS. Then for some constants c and C depending only on c_j and C_j for $j = 1, ..., 5$,*

$$\inf_{w \in \mathcal{G}_\rho} P_w\left(\widehat{T} > c^P_{1-\alpha}\right) \geqslant 1 - Cn^{-c} \text{ for all } n. \tag{2.11}$$

2.4.4 Consistency Against Local One-Dimensional Alternatives

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

Theorem 10. *Let $P = PIA$ or RMS. Assume that $a_n\sqrt{n/\log n} \to \infty$. Then*

$$P_{w_n}\left(\widehat{T} > c_{1-\alpha}^{P}\right) \to 1 \text{ as } n \to \infty. \tag{2.12}$$

Comment 7. *Recall the CMI model from the first example mentioned in the Introduction where $m(X, W, \theta) = \theta\tilde{m}(X, W)$ and $E[\tilde{m}(X, W)|X] > 0$ a.s. The theorem above shows that the test developed in this paper is consistent against sequences of alternatives $\theta_0 = \theta_{0,n}$ whenever $\theta_{0,n}\sqrt{n/\log n} \to \infty$ in this model whereas the test of [5] is consistent whenever $\theta_{0,n}\sqrt{n} \to \infty$. Hence, my test is consistent against nearly the same sequence of alternatives in this model as the test of [5]. The additional $\sqrt{\log n}$ factor is the cost for having higher power in other classes of models.*

2.4.5 Uniform Consistency Against Holder Smoothness Classes

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

$w \in \mathcal{G}$ and corresponding $f = f(w)$, let

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

denote the distance between w and H_0 over the set S_{ϑ}. For the next theorem, I will use ρ_{ϑ}-metric (instead of ρ-metric) to measure the distance between alternatives and the null hypothesis. Such restrictions are quite common in the literature. See, for example, [42] and [74]. Let $\mathcal{G}_{\vartheta}$ be the subset of all elements of $\mathcal{G}$ such that $\inf_{w \in \mathcal{G}_{\vartheta}} \rho_{\vartheta}(w, H_0)/h_{\min}^{\zeta} \to \infty$ if $\zeta \leqslant d$ and $\inf_{w \in \mathcal{G}_{\vartheta}} \rho_{\vartheta}(w, H_0)(n/\log n)^{\zeta/(2\zeta+d)} \to \infty$ if $\zeta > d$. Then

Theorem 11. *Let $P = PIA$ or RMS. Then*

$$\inf_{w \in \mathcal{G}_{\vartheta}} P_w \left(\widehat{T} > c_{1-\alpha}^{P} \right) \to 1 \text{ as } n \to \infty. \tag{2.13}$$

Comment 8. *Recall the CMI model from the second example mentioned in the Introduction where $m(X, W, \theta) = \tilde{m}(X, W) + \theta$. Assume that $X \in \mathbb{R}$ and $E[\tilde{m}(X, W)|X] = -|X|^{\nu}$ with $\nu > 1$. In this model, the identified set is $\Theta_I = \{\theta \in \mathbb{R} : \theta \leqslant 0\}$. The theorem above shows that the test developed in this paper is consistent against sequences of alternatives $\theta_0 = \theta_{0,n}$ whenever $\theta_{0,n}(n/\log n)^{\nu/(2\nu+1)} \to \infty$. At the same time, it follows from [10] that the test of [5] is consistent only if $\theta_{n,0} n^{\nu/(2\nu+2)} \to \infty$, so their test has a slower rate of consistency than that developed in this paper by a polynomial order.*

2.4.6 Lower Bound on the Minimax Rate of Testing

In this section, I give a lower bound on the minimax rate of testing. For any $X = \{X_i\}_{i=1}^{\infty}$ satisfying A13, let $\mathcal{G}_X$ denote the set of all models w in $\mathcal{G}$ with the sequence of design points X. For given X and S_{ϑ} defined in the previous section, let $N(h, S_{\vartheta})$ be the largest m such that there exists $\{x_1, ..., x_m\} \subset S_{\vartheta}$ with $\|x_i - x_j\| \geqslant h$ for all $i, j = 1, ..., m$ if $i \neq j$. I will assume that $N(h, S_{\vartheta}) \geqslant Ch^{-d}$ for all $h \in (0, 1)$ and sufficiently large n for some constant $C > 0$. In an iid setting, this condition holds a.s. under the conditions of Lemma 18. Let $\phi_n(Y_1, ..., Y_n)$, $n \geqslant 1$, denote a sequence of tests. In other words, $\phi_n(Y_1, ..., Y_n)$ denotes the probability of rejecting the null hypothesis upon observing sample $Y = (Y_1, ..., Y_n)$.

Theorem 12. *Assume that (i)* $N(h, S_\vartheta) \geqslant Ch^{-d}$ *for all* $h \in (0, 1)$*, sufficiently large* n*, and some* $C > 0$*, (ii)* $\varsigma = [\tau]$*, and (iii)* $r_n(n/\log n)^{\tau/(2\tau+d)} \to 0$ *as* $n \to \infty$ *for some sequence of positive numbers* r_n*. Then for any sequence of tests* $\phi_n(Y_1, ..., Y_n)$ *with* $\sup_{w\in\mathcal{G}_0\cap\mathcal{G}_X} E_w[\phi_n(Y_1, ..., Y_n)] \leqslant \alpha$*,*

$$\inf_{w\in\mathcal{G}_X,\rho_\vartheta(w,H_0)\geqslant Cr_n} E_w[\phi_n(Y_1, ..., Y_n)] \leqslant \alpha + o(1) \text{ as } n \to \infty. \tag{2.14}$$

Comment 9. *Since* $\mathcal{F}_{[\tau]}(\tau, L) \subset \mathcal{F}(\tau, L)$*, the same lower bound applies for the class* $\mathcal{F}(\tau, L)$ *as well. The same lower bound also applies with* $\mathcal{G}$ *instead of* $\mathcal{G}_X$*. Comparing this result with that in Theorem 4 shows that the test presented in this paper is minimax rate optimal (for almost all sequences* $\{X_i\}_{i=1}^{\infty}$*) if* $\zeta = \tau > d$*. The lower bound is not achieved when* $\zeta < \tau$ *or* $\tau \leqslant d$*. Two possibilities arise in these cases: the lower bound is not tight or the test is not minimax rate optimal. When* $\zeta < \tau$*, it is easy to see that the lower bound is achieved by the test with a higher order kernel and an appropriate (smoothness dependent) bandwidth value. Hence, the lower bound in this case is tight, and the test is not minimax rate optimal. When* $\zeta = \tau \leqslant d$*, the test does not achieve the lower bound because of the constraint on* $h_{\min}$ *imposed in A16. It is unknown whether this constraint can be relaxed without imposing existence of higher moments of* ε_i*'s beyond those imposed in A14(ii). [12] show, however, that if one assumes existence of finite moment generating function of* ε_i*'s, then one can construct a test that will achieve the rate in the lower bound derived in Theorem 12. Actually, under the assumption of finite moment generating function, the same rate can be obtained by the test considered in this paper. Specifically, assume in A14 that* $E[\exp(|\varepsilon_{i,m}|/C_2)] \leqslant C_2$ *for all* $i = 1, ..., n$ *and* $m = 1, ..., p$ *instead of* $E[\max_{m=1,...,p}|\varepsilon_{i,m}^4|] \leqslant C_2$ *for all* $i = 1, ..., n$ *and assume in A16 that* $h_{\min} = C_3(\log n)^8/n^d$ *instead of* $h_{\min} = C_3(\log n/n)^{1/(3d)}$*. Then Theorem 8 continues to hold (with the only difference in the proof that now one applies Corollary 2.3, case*

E.4 in [33] instead of case E.5 in Lemma 25) and the argument like that in the proof of Theorem 11 shows that the test is uniformly consistent against the set of models $\mathcal{G}_\vartheta$ if $\inf_{w\in\mathcal{G}_\vartheta}\rho_\vartheta(w,H_0)(n/\log n)^{\zeta/(2\zeta+d)}\to\infty$. This implies that the test is rate optimal when $\zeta=\tau\leqslant d$.

2.5 Extentions

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

Infinitely Many CMI. In many cases the parameter θ is restricted by a countably infinite number of CMI, i.e. $p=\infty$. For example, recall the English auction model and the model with interval data from Section 2.2. In those models, inequalities (2.3) and (2.4)-(2.5) hold for all $v\in\mathbb{R}$. Taking rational values of v leads to a countably infinite number of CMI. Note that the last step does not change the identified set if left-hand sides of these inequalities are continuous in v or, at least, right or left continuous.

Let $\tilde{m}:\mathbb{R}^d\times\mathbb{R}^k\times\Theta\to\mathbb{R}^{\mathbb{N}}$ be some known function where $\mathbb{N}$ denotes the set of natural numbers. Suppose that $\theta\in\Theta$ satisfies

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

Given θ_0, define $\tilde{f}(X)=E[\tilde{m}(X,W,\theta_0)|X]$. In addition, denote $\tilde{\varepsilon}_i=\tilde{m}(X_i,W_i,\theta_0)-\tilde{f}(X_i)$, and $\tilde{\Sigma}_i=E[\tilde{\varepsilon}_i\tilde{\varepsilon}_i^T|X_i]$. Let $\{p_n\}_{n=1}^\infty$ be a sequence of natural numbers converging to infinity. Consider the test based on the first $p=p_n$ inequalities. More precisely, let $m:\mathbb{R}^d\times\mathbb{R}^k\times\Theta\to\mathbb{R}^p$ be the vector-valued function whose j-th component coincides with j-th component of $\tilde{m}$ for all $j=1,...,p$, and consider the test described in Section 2.3 based on inequalities $E[m(X,W,\theta)|X]\leqslant 0$ a.s. Denote its critical value by $c_{1-\alpha}^P$ with $P=PIA$ or RMS. In addition, let me use all the notation defined in

Sections 2.3 and 2.4 and corresponding to the test based on the function m.

Let $\mathcal{G}$ denote the set of all models satisfying A13-A19 for all n where $p = p_n$ is now understood to be a function of n and where ε_i, Σ_i, f, and p in A14 and A15 are replaced by $\tilde{\varepsilon}_i$, $\tilde{\Sigma}_i$, $\tilde{f}$, and ∞, respectively. Let $\mathcal{G}_0$ denote the set of models in $\mathcal{G}$ satisfying $\tilde{f} \leqslant 0$ a.s., $\mathcal{G}_{00}$ denote the set of models in $\mathcal{G}_0$ satisfying $\tilde{f} = 0$ a.s., and $\mathcal{G}_\rho$ denote the set of models in $\mathcal{G}$ satisfying $\rho(w, H_0) \geqslant \rho$ where $\rho(w, H_0)$ is defined as in (2.10) with $\tilde{f}$ and ∞ instead of f and p, respectively. An advantage of the finite sample approach used in this paper is that it immediately gives certain conditions that ensure that such a test maintains the required size as $n \to \infty$.

Corollary 1. *Let $P = PIA$ or RMS. Assume that $p_n \log n \leqslant C_6 n^{c_6}$ for some sufficiently small and large constants c_6 and C_6, respectively.*[13]*. Then for some constants c and C depending only on c_j and C_j for $j = 1, ..., 6$,*

$$\sup_{w \in \mathcal{G}_0} P_w \left(\widehat{T} > c^P_{1-\alpha} \right) \leqslant \alpha + Cn^{-c} \text{ for all } n. \tag{2.15}$$

In addition,

$$\inf_{w \in \mathcal{G}_{00}} P_w \left(\widehat{T} > c^P_{1-\alpha} \right) \geqslant \alpha - Cn^{-c} \text{ for all } n. \tag{2.16}$$

Finally,

$$\inf_{w \in \mathcal{G}_\rho} P_w \left(\widehat{T} > c^P_{1-\alpha} \right) \geqslant 1 - Cn^{-c} \text{ for all } n. \tag{2.17}$$

Comment 10. *This corollary shows that the test has the correct asymptotic size, is asymptotically not conservative, and is consistent against fixed alternatives outside of the set Θ_I.*

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

$$E[m(X, W, \theta) | X^1, X^2 = x_0] \leqslant 0 \, a.s. \tag{2.18}$$

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

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

dimension of X is $d = d_1 + d_2$). CMI of the form (2.18) arise in nonparametric and semiparametric inference. For example, recall the English auction model from Section 2.2. In that model, suppose that the set of covariates is $X = (X^1, X^2)$ so that $F = F(v, X^1, X^2)$. Suppose that the point $X^2 = x_0$ is of interest. Denote $\tilde{F}(v, X^1) = F(v, X^1, x_0)$. Then inequality (2.3) leads to

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

Parameterizing the function $\tilde{F}(\cdot,\cdot)$ gives inequalities of the form (2.18). Note that parameterizing $\tilde{F}(\cdot,\cdot)$ instead of $F(\cdot,\cdot,\cdot)$ reduces the risk of misspecification, which makes this approach attractive when the only interesting value of X^2 is x_0.

As above, given θ_0, define $f(X) = E[m(X, W, \theta_0)|X]$. In addition, denote $\varepsilon_i = m(X_i, W_i, \theta_0) - f(X_i)$, and $\Sigma_i = E[\varepsilon_i(\varepsilon_i)^T|X_i]$. Let $\widehat{\Sigma}_i$ be an estimator of Σ_i ($i = 1, ..., n$) as described in Section 2.3.3. Let N be a subset of all observations $i = 1, ..., n$ such that $\|X_i^2 - x_0\| < a$ for all $i \in N$. It will be assumed that $a = a_n \to 0$ as $n \to \infty$. Denote the number of elements in N by n_a. Without loss of generality, I assume that observations in N are those corresponding to $i = 1, ..., n_a$. In order to test inequalities (2.18), consider the test described in Section 2.3 based on the data $\{(X_i, W_i)\}_{i \in N}$. Denote its test statistic by $\widehat{T}$ and its critical value by $c_{1-\alpha}^P$ with $P = PIA$ or RMS.

Let $\mathcal{G}$ denote the set of models satisfying A13-A19 for all n with n replaced by n_a, with d replaced by d_1 in A13 and A16, and such that for all these models, $|f_m(X)| \leqslant C_6 a_n$ for all $i \in N$ and $m = 1, ..., p$, and $a_n\sqrt{n_a h_{\max}^{d_1} \log n} \leqslant C_6 n^{-c_6}$ for sufficiently small and large constants c_6 and C_6, respectively. Let $\mathcal{G}_0$ denote the set of all models in $\mathcal{G}$ satisfying $f(X) \leqslant 0$ a.s., $\mathcal{G}_{00}$ denote the set of models in $\mathcal{G}_0$ satisfying $f(X) = 0$ a.s. Denote $\mathcal{N}_a = \{(i, m) : i = 1, ..., \infty, m = 1, ..., p, \|X_i^2 - x_0\| \leqslant a\}$. Define the distance between the model $w \in \mathcal{G}$ and the null hypothesis by

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

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

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

$$\sup_{w \in \mathcal{G}_0} P_w \left(\widehat{T} > c^P_{1-\alpha} \right) \leqslant \alpha + Cn^{-c} \text{ for all } n. \tag{2.19}$$

In addition,

$$\inf_{w \in \mathcal{G}_{00}} P_w \left(\widehat{T} > c^P_{1-\alpha} \right) \geqslant \alpha - Cn^{-c} \text{ for all } n. \tag{2.20}$$

Finally,

$$\inf_{w \in \mathcal{G}_\rho} P_w \left(\widehat{T} > c^P_{1-\alpha} \right) \geqslant 1 - Cn^{-c} \text{ for all } n. \tag{2.21}$$

Comment 11. *(i) Note that in an iid setting, if $f(x^1, x_0) > 0$ for some x^1 such that (x^1, x_0) is inside of the support of X, then it follows as in the proof of Lemma 18 that $\rho(w, H_0) > 0$ a.s. So, the corollary above shows that the test has correct asymptotic size, is asymptotically not conservative, and is consistent against any fixed alternative outside of the set Θ_I.*

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

(iii) Condition $a_n \sqrt{n_a h_{\max}^{d_1} \log n} \leqslant C_6 n^{-c_6}$ in this corollary is required to ensure that the bias due to using data with $X_i^2 \neq x_0$ is asymptotically negligible. Given that small values of a_n lead to small effective sample size n_a while small values of $h_{\max}$ lead to large variance of the kernel estimator, it is useful to set $h_{\max} \to 0$ as $n \to \infty$ to balance these effects.

2.6 Monte Carlo Results

In this section, I present results of two Monte Carlo simulation studies. The aim of these simulations is twofold. First, I demonstrate that my test accurately maintains size in finite samples. Second, I compare relative advantages and disadvantages of my test and the tests of [5], [36], and [74]. The methods of [5] and [74] are most appropriate for detecting flat alternatives, which represent one-dimensional local alternatives. These methods have low power against alternatives with peaks, however. The test of [36] has higher power against latter alternatives, but it requires knowing smoothness properties of the moment functions. The authors suggest certain rule-of-thumb techniques to choose a bandwidth value. Finally, the main advantage of my test is its adaptiveness. In comparison with [5] and [74], my test has higher power against alternatives with peaks. In comparison with [36], my test has higher power when their rule-of-thumb techniques lead to an inappropriate bandwidth value.[14] For example, this happens when the underlying moment function is mostly flat but varies significantly in the region where the null hypothesis is violated (the case of spatially inhomogeneous alternatives, see [76]).

First simulation study. The data generating process is

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

where X_i's are equidistant on the $[-2, +2]$ interval[15], Y_i's and ε_i's are scalar random variables, and L, M, and m are some constants. Depending on the experiment, ε_i's have either normal or (continuous) uniform distribution with mean zero. In both cases, the variance of ε_i's is 0.01. I consider the following specifications for parameters. Case 1: $L = M = m = 0$. Case 2: $L = 0.1$, $M = 0.2$, $m = 0.02$. Case 3: $L = M = 0$, $m = -0.02$. Case 4: $L = 2$, $M = 0.2$, $m = 0.2$. Note that $E[Y|X] \leqslant 0$ a.s. in cases 1 and 2 while $P(E[Y|X] > 0) > 0$ in cases 3 and 4. In case 3, the alternative is flat. In case 4, the alternative has a peak in the region where the null hypothesis is violated. I have chosen parameters so that rejection probabilities are strictly greater than 0 and strictly smaller than 1 in most cases so that meaningful comparisons are possible. I generate samples $(X_i, Y_i)_{i=1}^n$ of size $n = 250$ and 500. In all cases, I consider tests with the nominal size 10%. The results are based on 1000 simulations for each specification.

For the test of [5], I consider their Kolmogorov-Smirnov test statistic with boxes and truncation parameter 0.05. I simulate both plugin (AS, plugin) and GMS (AS, GMS) critical values based on the asymptotic approximation suggested in their paper.

[14] When their rule-of-thumb works well and moment functions are sufficiently smooth, the test of [36] often yielded the best results in my simulations.

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

All other tuning parameters are set as prescribed in their paper.

Implementing all other tests requires selecting a kernel function. In all cases, I use[16]

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

For the test of [36], I use their kernel type test statistic with critical values based on the multiplier bootstrap both with (CLR, $\widehat{V}$) and without (CLR, V) the set estimation. Both [36] and [74] (LSW) circumvent edge effects of kernel estimators by restricting their test statistics to the proper subsets of the support of X. To accomodate this, I select the 10%th and 90%th percentiles of the empirical distribution of X as bounds for the set over which the test statistics are calculated. Both tests are nonadaptive. In particular, there is no formal theory on how to choose bandwidth values in their tests, so I follow their informal suggestions. For the test of [74], I use their test statistic based on one-sided L_1-norm.

Parameters for the test developed in this paper are chosen according to recommendations in Section 2.3.1. Specifically, the largest bandwidth value, $h_{\max}$, is set to be equal to the length of the support of the empirical distribution.[17] The smallest bandwidth value, $h_{\min}$, is set as $h_{\min} = 0.2h_{\max}(\log n/n)^{1/3}$. The scaling parameter, a, equals 0.5 so that the set of bandwidth values is

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

I estimate Σ_i using the method of [97]. Specifically, I rearrange the data so that $X_1 \leqslant ... \leqslant X_n$ and set $\widehat{\Sigma}_i = \widehat{\Sigma} = \sum_{i=2}^n (Y_i - Y_{i-1})^2/(2n)$. Finally, for the RMS critical value, I set $\gamma = 0.1/\log(n)$ to make meaningful comparisons with the test of [36]. In

[16] This kernel function does not coincide with recommendations in Section 2.3.1 (where I recommended the kernel $K(x) = 0.75(1 - x^2)_+$). I use this kernel function because it was used in other simulation studies; see, in particular [74]. Note, however, that for the test statistic in this paper, multiplicative constant in the kernel function has no effect (it cancells out because of studentization), and so using kernels $K(x) = 1.5(1 - 4x^2)_+$ and $K(x) = 0.75(1 - x^2)_+$ gives numerically the same values of the test statistic if all bandwidth values for the former kernel are twice as large as bandwidth values for the latter kernel.

[17] In Section 2.3.1, I recommend setting the largest bandwidth value as one half of the length of the support of the empirical distribution. The difference is explained by different scaling of the kernel function.

Table 2.1: Results of Monte Carlo Experiments, $n = 250$

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

all bootstrap procedures, for all tests, I use 500 repetitions.

The results of the first simulation study are presented in table 1 for $n = 250$ and in table 2 for $n = 500$. In both tables, my test is denoted as Adaptive test with plug-in and RMS critical values. Consider first results for $n = 250$. In case 1, where the null hypothesis holds, all tests have rejection probabilities close to the nominal size 10% both for normal and uniform disturbances. In case 2, where the null hypothesis holds but the underlying regression function is mainly strictly below the borderline, all tests are conservative. When the null hypothesis is violated with a flat alternative (case 3), the tests of [5] and [74] have highest rejection probabilities as expected from the theory. In this case, my test is less powerful in comparison with these tests and somewhat similar to the method of [36]. This is compensated in case 4 where the null hypothesis is violated with the peak-shaped alternative. In this case, the power of my test is much higher than that of competing tests. This is especially true for my test with RMS critical values whose rejection probability exceeds 80% while rejection probabilities of competing tests do not exceed 20%. Note that all results are stable across distributions of disturbances. Also note that my test with RMS critical values has higher power than the test with plugin critical values in case 4. So, among these two tests, I recommend the test with RMS critical values. Results for $n = 500$ indicate a similar pattern.

Second simulation study. In the second simulation study, I compare the power function of the test developed in this paper with that of the Andrews and Shi's (2013)

Table 2.2: Results of Monte Carlo Experiments, $n = 500$

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

test, which is most closely related to my method. For my test, I use the RMS critical value. For the test of [5], I use their GMS critical value. The data generating process is

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

where X_i's are again equidistant on the $[-2, +2]$ interval, Y_i's and ε_i's are scalar random variables, m and τ are some constants, and $\phi(\cdot)$ is the pdf of the standard Gaussian distribution. In this experiment, ε_i's have $N(0,1)$ distrubution. I use samples $(X_i, Y_i)_{i=1}^n$ of size $n = 250$. Both tests are based on the same specifications as in the first simulation study except that now I use 100 repetitions for all simulation procedures in order to conserve computing time. At each point, the rejection probabilities are estimated using 500 simulations.

Note that τ is naturally bounded from below because τ and $-\tau$ yield the same results. So, I set $\tau \geqslant 0$. In addition, $E[Y|X] \leqslant 0$ a.s. if $m \leqslant -1$. Therefore, I set $m \geqslant -1$. Figure 1 shows the difference between the rejection probabilities of my test and of the test of [5]. This figure shows that the rejection probability of the test developed in this paper is higher than that of the test of [5] in most cases and is strictly higher over a wide region of parameter values. The exception is a narrow region where τ is close to 0 (flat alternatives) and m is close to -1. Concluding this section, I note that all simulation results are consistent with the presented theory.

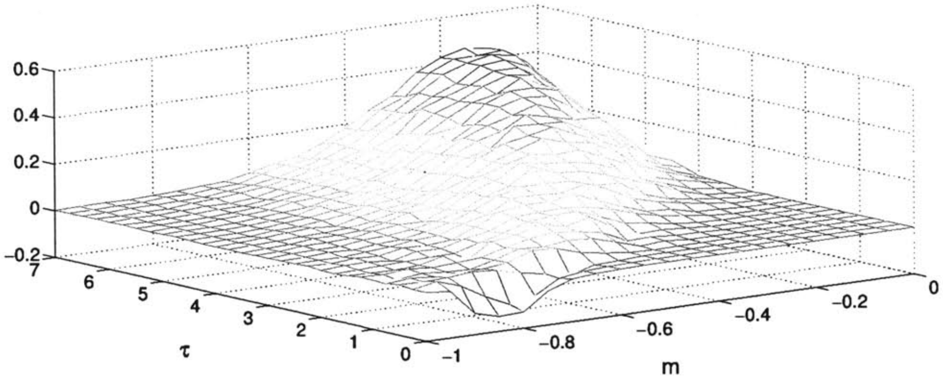


Figure 2-1: The difference between the rejection probabilities of the test developed in this paper and of the test of Andrews and Shi (2013) (with RMS and GMS critical values correspondingly). The nominal size is 10%. Results are based on 500 simulations. The figure shows that the rejection probability of the test developed in this paper is higher than that of the test of Andrews and Shi (2013) in most cases and is strictly higher over a wide region of parameter values.

2.7 Conclusions

In this paper, I develop a new test of conditional moment inequalities. In contrast to some other tests in the literature, my test is directed against general nonparametric alternatives yielding high power in a large class of CMI models. Considering kernel estimates of moment functions with many different values of the bandwidth parameter allows me to construct a test that automatically adapts to the unknown smoothness of moment functions and selects the most appropriate testing bandwidth value. The test developed in this paper has uniformly correct asymptotic size, no matter whether the model is identified, weakly identified, or not identified, is consistent against any fixed alternative outside of the set Θ_I, and is uniformly consistent against certain, but not all, large classes of smooth alternatives whose distance from the null hypothesis converges to zero at a fastest possible rate. The tests of [5] and [74] have nontrivial power against $n^{-1/2}$-local one-dimensional alternatives whereas my method only allows for nontrivial testing against $(n/\log n)^{-1/2}$-local alternatives of this type. The additional $(\log n)^{1/2}$ factor should be regarded as the price for having fast rate of uniform consistency. There exist sequences of local alternatives against which their tests are not consistent whereas mine is. Monte Carlo experiments give an example of a CMI model where finite sample power of my test greatly exceeds that of competing tests.

2.8 Appendix. Proofs

This Appendix contains proofs of all results stated in the main part of the paper. Section 2.8.1 gives a proof of the uniform consistency of the estimator $\widehat{\Sigma}_i$ of Σ_i described in Section 2.3.3. I provide the proof because I was not able to find it in the literature. Section 2.8.2 derives a bound on the modulus of continuity in the spectral norm of the square root operator on the space of symmetric positive semidefinite matrices. Section 2.8.3 gives sufficient conditions for A13 in the main part of the paper. Section 2.8.4 explains an anticoncentration inequality for the maximum of Gaussian random variables with unit variance. Section 2.8.5 describes a result on Gaussian random variables that is used in the proof of the lower bound on the minimax rate. Section 2.8.6 develops some preliminary technical results necessary for the proofs of the main theorems. Finally, Section 2.8.7 presents the proofs of the theorems stated in the main part of the paper.

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

2.8.1 Lemma on the Estimator of Σ_i

Lemma 15. *Let $\widehat{\Sigma}_i$ be an estimator of Σ_i described in Section 2.3.3. Let A13-A15 hold. In addition, assume that (i) $E[|\varepsilon_{i,m}|^{4+\delta}] \leqslant C$ for all $i = 1, ..., n$ and $m = 1, ..., p$, (ii) $b \leqslant Cn^{-c}$, (iii) $\min_{i=1,...,n} |J(i)|/n^{1/(2+\delta)} \geqslant cn^C$, (iv) $\|\Sigma_i - \Sigma_j\| \leqslant C\|X_i - X_j\|$. Then A18 holds.*

Comment 12. *Note that under assumptions of Lemma 18, which is described below, condition (iii) above follows from $n^{(1+\delta)/(2+\delta)}b^d \geqslant cn^C$, which is an elementary condition.*

Proof. By definition,

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

Since all norms on the finite-dimensional linear space are equivalent (Theorem 1.6 in [70]), it is enough to prove that

$$P(\max_{i=1,...,n} |\widehat{\Sigma}_{i,m_1 m_2} - \Sigma_{i,m_1 m_2}| > Cn^{-c}) \leqslant Cn^{-c}$$

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

Let $M = n^{1/(4+\delta/2)}$. Consider a truncated version of ε_i's: $\tilde{\varepsilon}_i = \varepsilon_i I\{\varepsilon_i \leqslant M\}$. Since $E[|\varepsilon_i|^{4+\delta}] \leqslant C$, it follows that $E[\max_{i=1,...,n} |\varepsilon_i|] \leqslant Cn^{1/(4+\delta)}$ (see Lemma 2.2.2 in [111]). Then Markov inequality gives

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

So,

$$P_1 = P(\max_{i=1,...,n} |\tilde{\varepsilon}_i - \varepsilon_i| > 0) \leqslant Cn^{-c}.$$

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

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

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

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

Then

$$P(\max_{i=1,\dots,n} |\bar{\Sigma}_i - \widehat{\Sigma}_i| > 0) = P_1 \leqslant Cn^{-c}.$$

Therefore, for sufficiently small c and sufficiently large C,

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

for all n. By the union bound,

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

Further,

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

where

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

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

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

By A15, $|f(X_{k(j)}) - f(X_j)| \leqslant L\|X_{k(j)} - X_j\| \leqslant 2Lb$. Since b converges to zero at a polynomial rate, $P_1 = 0$ if c and C in the definition of P_1 are sufficiently small and large, respectively. Consider P_3. Note that $P_3 \leqslant P_{31} + P_{32}$ where

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

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

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

Then Hoeffding inequality gives (see proposition 1.3.5 in [41])

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

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

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

2.8.2 Continuity of the Square Root Operator on the Set of Positive Semidefinite Matrices

Lemma 16. *Let A and B be $p \times p$-dimensional symmetric positive semidefinite matrices. Then $\|A^{1/2} - B^{1/2}\| \leqslant p^{1/2}\|A - B\|^{1/2}$.*

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

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

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

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

For $P = A, B$, $P^{1/2}$ has the same eigenvectors as P with corresponding eigenvalues equal to $\lambda_1^{1/2}(P), ..., \lambda_n^{1/2}(P)$. Therefore, for any $i = 1, ..., p$,

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

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

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

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

2.8.3 Primitive Conditions for A1

In this section, I give a counter-example for the statement that for A13 to hold, it suffices to assume that X_i's are sampled from a distribution that is absolutely continuous with respect to Lebesgue measure, has bounded support, and whose density is bounded from above and away from zero on the support. I also prove that A13 holds if, in addition to above conditions, one assumes that the support is a convex set.

Lemma 17. *There exists a probability distribution on $\mathbb{R}^2$ with bounded support such that this distribution is uniform on its support and if X_i's are sampled from this distribution, then A13 fails.*

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

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

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

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

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

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

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

Now I give a sufficient primitive condition for A13.

Lemma 18. *Suppose that A16 holds. If X_i's are sampled from a distribution that is absolutely continuous with respect to Lebesgue measure, has bounded and convex support $S \subset \mathbb{R}^d$, and whose density is bounded from above and away from zero on the support, then A13 holds for sufficiently large n a.s.*

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

Note that δ-covering numbers of the set S satisfy $N(\delta) \leqslant C/\delta^d$. Consider the lower bound in A13(ii). For each $h \in H_n$, consider the set of covering balls with centers $G_{h,1}$,...,$G_{h,N(h)}$ and radii $\delta_h = h/2$. Then for each X_i and $h \in H_n$, there exists some $j \in \{1, ..., N(h)\}$ such that $B(X_i, h) \supset B(G_{h,j}, \delta_h)$. Thus, it is enough to prove the lower bound for the number of observations dropping into these covering balls. Since the density is bounded away from zero and from above, there exist some constants $c, C > 0$ such that for each $h \in H_n$ and $j = 1, ..., N(h)$, $ch^d \leqslant P(X_i \in B(G_{h,j}, \delta_h)) \leqslant Ch^d$. Denote $I_{h,j}(X_i) = I\{X_i \in B(G_{h,j}, \delta_h)\}$. Bernstein inequality (see proposition 1.3.2 in [41]) gives

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

Then by union bound and A16,

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

By A16, $nh_{\min}^d > Cn^c$. So, summing the probabilities above over n, I conclude, by the Borel-Cantelli lemma, that the lower bound in A13(ii) holds for sufficiently large n a.s. A similar argument gives the upper bound. So, A13 holds. □

2.8.4 Anticoncentration Inequality for the Maximum of Gaussian Random Variables

In this section, I describe an upper bound on the pdf of the maximum of correlated Gaussian random variables derived in [34]. Let $\{Z_i : i = 1, ..., S\}$ be a set of standard Gaussian (possibly correlated) random variables. Define $W = \max_{i=1,...,S} Z_i$ and let $f_W(\cdot)$ denote its pdf. Then

Lemma 19. $\sup_{w\in\mathbb{R}} f_W(w) \leqslant C\sqrt{\log S}$ *for some universal constant* C.

Proof. Theorem 3 in [34] proves that $\sup_{w\in\mathbb{R}} f_W(w) \leqslant CE[W]$. In addition, it follows from the same argument as in Lemma 22 that $E[W] \leqslant C\sqrt{\log S}$. Combining these bounds gives the result. □

2.8.5 Result on Gaussian Random Variables

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

Lemma 20. *Let ξ_n, $n = 1, ..., \infty$, be a sequence of independent standard Gaussian random variables and $w_{i,n}$, $i = 1, ..., n$, $n = 1, ..., \infty$, be a triangular array of positive numbers. If $w_{i,n} \leqslant C\sqrt{\log n}$ with $C \in (0, 1)$ for all $i = 1, ..., n$, $n = 1, ..., \infty$, then*

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

Proof. The proof is closely related to that in Lemma 6.2 in [42]. Denote $Z_{i,n} = \exp(w_{i,n}\xi_i - w_{i,n}^2/2)$ and $t_n = (E[(\sum_{i=1}^n Z_{i,n}/n - 1)^2])^{1/2}$. Note that $E[Z_{i,n}] = 1$ and $E[Z_{i,n}^2] = \exp(w_{i,n}^2)$. Thus,

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

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

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

The result follows. □

2.8.6 Preliminary Technical Results

In this section, I derive some necessary preliminary results that are used in the proofs of the theorems stated in the main part of the paper. It is assumed throughout that conditions A13-A19 hold. I will use the following additional notation. Let $\{\psi_n\}_{n=1}^{\infty}$ be a sequence of positive real numbers such that $\psi_n \geqslant C_\psi(p\log n)^{1/2}/n^{c_\psi}$ for some sufficiently large $C_\psi > 0$ and sufficiently small $c_\psi > 0$ and $\psi_n \leqslant Cn^{-c}$ for all n. For any $\lambda \in (0, 1)$, define $c_{1-\lambda}^{PIA,0} \in \mathbb{R}$ by analogy with $c_{1-\lambda}^{PIA}$ with Σ_i used instead of $\widehat{\Sigma}_i$ for all $i = 1, ..., n$. Denote $S_n^D = \{s \in S_n : f_s/V_s > -c^{PIA,0}_{1-\gamma_n-\psi_n}\}$. For any $\lambda \in (0,1)$, define $c^D_{1-\lambda} \in \mathbb{R}$ by analogy with $c^{RMS}_{1-\lambda}$ with S_n^D used instead of S_n^{RMS}. Let $\{\epsilon_i : i = 1, ..., n\}$ be an iid sequence of p-dimensional standard Gaussian random vectors that are independent of the data. Denote $\widehat{e}_j = \widehat{\Sigma}^{1/2}\epsilon_j$ and $e_j = \Sigma^{1/2}\epsilon_j$. Note that $\widehat{e}_j$ is equal in distribution to $\tilde{Y}_j$. Finally, denote

$$\varepsilon_{(i,m,h)} = \sum_{j=1}^n w_h(X_i, X_j)\varepsilon_{j,m} \quad \text{and} \quad f_{(i,m,h)} = \sum_{j=1}^n w_h(X_i, X_j) f_m(X_j),$$
$$e_{(i,m,h)} = \sum_{j=1}^n w_h(X_i, X_j) e_{j,m} \quad \text{and} \quad \widehat{e}_{(i,m,h)} = \sum_{j=1}^n w_h(X_i, X_j)\widehat{e}_{j,m},$$
$$T^{PIA} = \max_{s \in S_n}(\widehat{e}_s/\widehat{V}_s) \quad \text{and} \quad T^{PIA,0} = \max_{s \in S_n}(e_s/V_s).$$

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

I start with a result on bounds for weights and variances of the kernel estimator. The same result can be found in [63].

Lemma 21. *There exist constants $c, C > 0$ such that for any $i, j = 1, ..., n$, $m = 1, ..., p$, and $h \in H_n$,*

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

and

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

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

Proof. By A13 and A17, for any $i = 1, ..., n$ and $h \in H_n$,

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

and

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

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

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

By A14, since $\sum_{j=1}^{n} w_h(X_i, X_j) = 1$,

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

and

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

The claim of the lemma follows. □

Lemma 22. $E[\max_{s\in S_n} |e_s/V_s|] \leqslant C(\log n)^{1/2}$ *uniformly over the set of models* $\mathcal{G}$. *In particular,* $c_{1-\lambda}^{PIA,0} \leqslant C\sqrt{\log n}/\lambda$ *for all* $\lambda \in (0, 1)$ *uniformly over the set of models* $\mathcal{G}$. *In addition,* $P(\max_{s\in S_n} |e_s/V_s| > C\sqrt{\log n}) \leqslant Cn^{-c}$ *for sufficiently small and large constants* c *and* C*, respectively, uniformly over the set of models* $\mathcal{G}$.

Proof. For any $s \in S_n$, e_s/V_s is a standard Gaussian random variable. Denote $\psi = \exp(x^2) - 1$. Let $\|\cdot\|_\psi$ denote ψ-Orlicz norm. It is easy to check that $\|e_s/V_s\|_\psi < C < \infty$. So, by Lemma 2.2.2 in [111],

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

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

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

for any $\lambda \in (0,1)$. So, $c_{1-\lambda}^{PIA,0} \leqslant C\sqrt{\log n}/\lambda$. The third result follows from Borell inequality (see, for example, Proposition A.2.1 in [111]). □

Lemma 23. *$P(\max_{s\in S_n}|\widehat{V}_s/V_s - 1| > Cn^{-c}) \leqslant Cn^{-c}$ and $P(\max_{s\in S_n}|V_s/\widehat{V}_s - 1| > Cn^{-c}) \leqslant Cn^{-c}$ uniformly over the set of models $\mathcal{G}$.*

Proof. By A14, for any $(i,m,h) \in S_n$,

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

In addition,

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

So,

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

So,

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

by A18. Combining this result with inequality $|x-1| \leqslant |x^2-1|$, which holds for any $x > 0$, yields the first result of the lemma. The second result follows from the first one and the inequality $|1/x - 1| < 2|x-1|$, which holds for any $|x-1| < 1/2$. □

Lemma 24. *$P(c_{1-\lambda-\psi_n}^{PIA,0} > c_{1-\lambda}^{PIA}) \leqslant Cn^{-c}$ and $P(c_{1-\lambda+\psi_n}^{PIA,0} < c_{1-\lambda}^{PIA}) \leqslant Cn^{-c}$ uniformly over all $\lambda \in (0,1)$[18] and over the set of models $\mathcal{G}$ where ψ_n is defined in the beginning of this section ($\psi_n \geqslant C_\psi(p\log n)^{1/2}/n^{c_\psi}$ and $\psi_n \leqslant Cn^{-c}$).*

Proof. Denote

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

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

and

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

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

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

Let A denote the event $\{\max_{j=1,...,n} \|\widehat{\Sigma}_j - \Sigma_j\| < C_4 n^{-c_4}\}$. By A18, $P(A) \geqslant 1 - Cn^{-c}$ as $n \to \infty$. Thus, it is enough to show that $c^{PIA,0}_{1-\lambda-\psi_n} \leqslant c^{PIA}_{1-\lambda}$ and $c^{PIA,0}_{1-\lambda+\psi_n} \geqslant c^{PIA}_{1-\lambda}$ on A.

As in the proof of Lemma 23, $\max_{s\in S_n} |V_s/\widehat{V}_s - 1| \leqslant Cn^{-c}$ on A. Lemma 22 shows that $E[\max_{s\in S_n} |e_s/V_s|] \leqslant C\sqrt{\log n}$. So, Markov inequality gives for any $B > 0$, on A,

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

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

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

where the last line follows from Lemma 16. So, conditional on Y_1^n, on A,

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

is a mean-zero Gaussian random variable with variance bounded by $Cp^2 n^{-c}$ for any $(i, m, h) \in S_n$. In addition, on A, $\max_{s\in S_n} V_s/\widehat{V}_s \leqslant 2$ for sufficiently large n. Thus, Markov inequality and the argument like that used in Lemma 22 yield

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

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

Note that $T^{PIA,0}$ is the maximum over $|S_n|$ standard Gaussian random variables. Since $|S_n| \leqslant Cn^{\phi}$ for some $\phi > 0$, Lemma 19 gives $c^{PIA,0}_{1-\lambda-\psi_n/2} - c^{PIA,0}_{1-\lambda-\psi_n} \geqslant c\psi_n/(\log n)^{1/2}$, so that

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

Now the first part of the lemma follows from

$$\begin{aligned}
P(T^{PIA} \leqslant c^{PIA,0}_{1-\lambda-\psi_n}|Y^n_1) &\leqslant P(T^{PIA,0} - p_1 - p_2 \leqslant c^{PIA,0}_{1-\lambda-\psi_n}|Y^n_1) \\
&\leqslant P(T^{PIA,0} - C\sqrt{\log n}pn^{-c}B \leqslant c^{PIA,0}_{1-\lambda-\psi_n}|Y^n_1) + 2/B \\
&\leqslant P(T^{PIA,0} \leqslant c^{PIA,0}_{1-\lambda-\psi_n/2}|Y^n_1) + 2/B \\
&\leqslant 1 - \lambda - \psi_n/2 + 2/B \\
&\leqslant 1 - \lambda
\end{aligned}$$

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

Lemma 25. $|P(\max_{s\in S_n}(\varepsilon_s/V_s) \leqslant c^{PIA,0}_{1-\lambda}) - (1-\lambda)| \leqslant Cn^{-c}$ *and* $|P(-\max_{s\in S_n}(\varepsilon_s/V_s) \leqslant c^{PIA,0}_{1-\lambda}) - (1-\lambda)| \leqslant Cn^{-c}$ *uniformly over all* $\lambda \in (0,1)$ *and over the set of models* $\mathcal{G}$.

Proof. By Lemma 21 and A14, for any $(i,m,h) \in S_n$ and $j = 1,...,n$,

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

Therefore, both claims of the lemma follows by combining A16 and Corollary 2.3, case E.5 in [33]. □

Lemma 26. *For sufficiently small and large constants c and C, respectively,*

$$\begin{aligned}
P(\max_{s\in S_n}|\varepsilon_s/V_s| &> C\sqrt{\log n}) \leqslant Cn^{-c}, \\
P(\max_{s\in S_n}|\varepsilon_s/\widehat{V}_s| &> C\sqrt{\log n}) \leqslant Cn^{-c},
\end{aligned}$$

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

Proof. The result for $\max_{s\in S_n}|\varepsilon_s/V_s|$ follows from combining Lemmas 22 and 25. The second result follows by noting that

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

and that $P(\max_{s\in S_n}|V_s/\widehat{V}_s| \leqslant 1+Cn^{-c}) \geqslant 1-Cn^{-c}$ by Lemma 23. □

Lemma 27. $P(\max_{s\in S_n\backslash S_n^D}\widehat{f}_s/\widehat{V}_s > 0) \leqslant Cn^{-c}$ *uniformly over the set of models $\mathcal{G}$.*

Proof. By Lemma 25,

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

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

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

Noting that $\gamma_n+\psi_n \leqslant Cn^{-c}$, which holds by the definition of ψ_n and A19, yields the result. □

Lemma 28. $P(S_n^D \subset S_n^{RMS}) \geqslant 1-Cn^{-c}$ *uniformly over the set of models $\mathcal{G}$.*

Proof. By Lemma 24, $P(c^{PIA,0}_{1-\gamma_n-\psi_n} > c^{PIA}_{1-\gamma_n}) \leqslant Cn^{-c}$. In addition, for any $x \in (-1,1)$,

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

So,

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

By Lemma 22, $c^{PIA,0}_{1-\gamma_n-\psi_n} \leqslant C(\log n)^{1/2}/(\gamma_n+\psi_n)$. By Lemma 23, $P(|\max_{s\in S_n^D}(V_s/\widehat{V}_s)-1| \leqslant Cn^{-c}) \geqslant 1-Cn^{-c}$. So, with probability at least $1-Cn^{-c}$,

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

Take $\chi_n = C(\log n)n^{-c}/(\gamma_n+\psi_n)$. Then $\chi_n \leqslant Cn^{-c}$ by the choice of ψ_n (recall that the constant c_ψ in the definition of ψ_n is sufficiently small). By Lemma 19,

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

Therefore,

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

The result follows since $\gamma_n+\psi_n+\chi_n \leqslant Cn^{-c}$ by the definitions of ψ_n and χ_n and A19. □

Lemma 29. *$P(S_n^{RMS} = S_n) \geqslant 1-Cn^{-c}$ uniformly over the set of models $\mathcal{G}_{00}$.*

Proof. By Lemma 24, $P(c^{PIA,0}_{1-\gamma_n-\psi_n} > c^{PIA}_{1-\gamma_n}) \leqslant Cn^{-c}$. It follows from Lemma 23 that $(\max_{s\in S_n}(V_s/\widehat{V}_s) \leqslant 1+Cn^{-c}) \geqslant 1-Cn^{-c}$. If $f = 0_p$, then for any $s \in S_n$, $\widehat{f}_s = \varepsilon_s$.

So,

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

Combining these results with Lemma 25 yields

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

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

Lemma 30. *$P(c_{1-\alpha}^{PIA} > C\sqrt{\log n}) \leqslant Cn^{-c}$ and $P(c_{1-\alpha}^{RMS} > C\sqrt{\log n}) \leqslant Cn^{-c}$ for sufficiently small and large c and C, respectively, uniformly over the set of models $\mathcal{G}$.*

Proof. Since $S_n^{RMS} \subseteq S_n$, it follows that $c_{1-\alpha}^{RMS} \leqslant c_{1-\alpha}^{PIA}$. Therefore, the second claim follows from the first one. To prove the first claim, note that by Lemma 24, $P(c_{1-\alpha/2}^{PIA,0} < c_{1-\alpha}^{PIA}) \leqslant Cn^{-c}$. In addition $c_{1-\alpha/2}^{PIA,0} \leqslant C\sqrt{\log n}$ by Lemma 22. Combining these results yields the asserted claim. □

Lemma 31. *Let $\tau > 1$, $L > 0$, $x = (x_1, ..., x_d) \in \mathbb{R}^d$, $h = (h_1, ..., h_d) \in \mathbb{R}^d$, and $g \in \mathcal{F}_\varsigma(\tau, L)$ for some $\varsigma = 1, ..., [\tau]$. Then $\partial g(x_1, ..., x_d)/\partial x_m \geqslant 0$ for all $m = 1, ..., d$ implies that for any $y = (y_1, ..., y_d) \in \mathbb{R}^d$ satisfying $0 \leqslant y \leqslant h$,*

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

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

Proof. For any $y = (y_1, ..., y_d) \in \mathbb{R}^d$ satisfying $0 \leqslant y \leqslant h$, let $l = y/\|y\|$. Then

$g^{(1,l)}(x) \geqslant 0$. If $g^{(1,l)}(x+tl) \geqslant 0$ for all $t \in (0, \|y\|)$, the result is obvious. If $g^{(1,l)}(x+t_0 l) = 0$ for some $t_0 \in (0, \|y\|)$, then $g^{(k,l)}(x+t_0 l) = 0$ for all $k = 1, ..., \varsigma$. If $\varsigma = [\tau]$, then by Holder smoothness, $g^{([\tau],l)}(x+tl) \geqslant -(L(t-t_0))^{\tau-[\tau]}$. Integrating it $[\tau]$ times gives

$$g(x+y) - g(x) \geqslant -\frac{L^{\tau-[\tau]}}{\prod_{j=1,\dots,[\tau]}(\tau-[\tau]+j)}\|y\|^{\zeta} \tag{2.22}$$

since $\zeta = \tau$ in this case. If $\varsigma < [\tau]$, then $g^{(\varsigma,l)}(x+tl) \geqslant -L(t-t_0)$. Integrating it ς times gives the inequality similar to (2.22) with $\varsigma+1$, ς, and L instead of ζ, $[\tau]$, and $L^{\tau-[\tau]}$ correspondingly. The result follows by noting that $\|y\| \leqslant \|h\|$. □

2.8.7 Proofs of Theorems

Proof of Theorem 1. Consider any $w \in \mathcal{G}_0$. For any $s \in S_n$, $f_s \leqslant 0$ since the kernel K is positive by A17. By Lemma 24, $P(c^{PIA,0}_{1-\alpha-\psi_n} > c^{PIA}_{1-\alpha}) \leqslant Cn^{-c}$. By Lemma 23, $P(\max_{s\in S_n}(V_s/\widehat{V}_s) \leqslant 1 + Cn^{-c}) \geqslant 1 - Cn^{-c}$. So,

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

Let $\chi_n = C(\log n)n^{-c}$. Since $P(\max_{s\in S_n}|\varepsilon_s/V_s| > C\sqrt{\log n}) \leqslant Cn^{-c}$ for sufficiently small and large c and C, respectively, by Lemma 26, an application of Lemma 19 shows that the last expression is bounded from below by

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

Then $P(\widehat{T} \leqslant c^{PIA}_{1-\alpha}) \geqslant 1 - \alpha - Cn^{-c}$ follows from this bound and Lemma 25 since $\psi_n + \chi_n \leqslant Cn^{-c}$.

Now consider the RMS critical value. By Lemma 28, $P(c^D_{1-\alpha} > c^{RMS}_{1-\alpha}) \leqslant Cn^{-c}$. By Lemma 27, $P(\max_{s\in S_n\setminus S^D_n} \widehat{f}_s/\widehat{V}_s > 0) \leqslant Cn^{-c}$. So,

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

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

Next consider any $w \in \mathcal{G}_{00}$ so that $f = 0_p$. By Lemma 24, $P(c^{PIA,0}_{1-\alpha+\psi_n} < c^{PIA}_{1-\alpha}) \leqslant Cn^{-c}$. By Lemma 23, $P(\min_{s\in S_n}(V_s/\widehat{V}_s) \geqslant 1 - Cn^{-c}) \geqslant 1 - Cn^{-c}$. So,

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

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

For the RMS critical value, note that by Lemma 29, $P(S^{RMS}_n = S_n) \geqslant 1 - Cn^{-c}$ whenever $f = 0_p$. So,

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

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

Proof of Theorem 2. For any $w \in \mathcal{G}_\rho$, there exist $i \in \mathbb{N}$ and $m = 1, ..., p$ such that $f_m(X_i) \geqslant 3\rho/4$. By A15, there exists a ball $B_\delta(X_i)$ with center at X_i and radius δ

such that $f_m(X_j) \geqslant \rho/2$ for all $X_j \in B_\delta(X_i)$. Note that δ can be chosen independently of w. So, for some $N \in \mathbb{N}$ and any $n \geqslant N$, there exists a triple $s_n = (i_n, m, h_n) \in S_n$ with h_n bounded away from zero such that $f_m(X_j) \geqslant \rho/2$ for all $X_j \in B_{h_n}(X_{i_n})$. Hence, $f_{s_n} \geqslant \rho/2$. Lemma 21 gives $V_{s_n} \leqslant Cn^{-\phi}$ for some $\phi > 0$, so $f_{s_n}/V_{s_n} > cn^{\phi}$. By Lemma 23, $P(|\widehat{V}_{s_n}/V_{s_n} - 1| > Cn^{-c}) \leqslant Cn^{-c}$. So, $P\{f_{s_n}/\widehat{V}_{s_n} > cn^{\phi}\} \geqslant 1 - Cn^{-c}$ for sufficiently small $c > 0$. Thus,

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

The result follows by noting that from Lemmas 26 and 30, $P(c^P_{1-\alpha} + \max_{s \in S_n} |\varepsilon_s/\widehat{V}_s| > C\sqrt{\log n}) \leqslant Cn^{-c}$. □

Proof of Theorem 3. Let $f^0 = f(w_0)$ and for all $n \geqslant 1$, $f^n = f(w_n)$. As in the proof of Theorem 2, since $\rho(w_0, H_0) > 0$, there exists $i \in \mathbb{N}$ such that $f^0_m(X_i) \geqslant 3\rho/4$ for some $m = 1, ..., p$ and $\rho > 0$. In addition, by A15, there exists a ball $B_\delta(X_i)$ such that $f^0_m(X_j) \geqslant \rho/2$ for all $X_j \in B_\delta(X_i)$. So, for some $N \in \mathbb{N}$ and any $n \geqslant N$, there exists a triple $s_n = (i_n, m, h_n) \in S_n$ with h_n bounded away from zero such that $f^0_m(X_j) \geqslant \rho/2$ for all $X_j \in B_{h_n}(X_{i_n})$. Hence, $f^n_{s_n} \geqslant a_n\rho/2$. By Lemma 21, $V_{s_n} \leqslant C/\sqrt{n}$. Then Lemma 23 gives $P(f^n_{s_n}/\widehat{V}_{s_n} > ca_n/\sqrt{n}) \to 1$. The same argument as in the proof of Theorem 2 yields

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

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

Proof of Theorem 4. First, consider $\tau \leqslant 1$ case. In this case, $\zeta = \tau$. Since $d \geqslant 1$, I have $\zeta \leqslant d$. Consider any $w \in \mathcal{G}_\vartheta$. Since $\inf_{w \in \mathcal{G}_\vartheta} \rho_\vartheta(w, H_0)/h^\zeta_{\min} \to \infty$, there exists a sequence a_n of positive numbers such that $a_n \to \infty$ and $\rho_\vartheta(w, H_0) > a_n h^\zeta_{\min}$, and so there exist $i \in \mathbb{N}_\vartheta$ and $m = 1, ..., p$ such that $f_m(X_i) \geqslant a_n h^\zeta_{\min}$. Let $s_n(w) = (i, m, h_{\min}) \in S_n$. By A15, $f_m(X_l) \geqslant ca_n h^\zeta_{\min}$ for all $l = 1, ..., n$ such that $X_l \in$

$B_{h_{\min}}(X_i)$. So, $f_{s_n(w)} \geqslant ca_n h_{\min}^{\zeta}$. By A16, $nh_{\min}^{3d}/\log n \geqslant c$. By Lemma 21, $V_{s_n(w)} \leqslant C/\sqrt{nh_{\min}^{d}}$. So,

$$f_{s_n(w)}/(V_{s_n(w)}\sqrt{\log n}) \geqslant ca_n\sqrt{nh_{\min}^{2\zeta+d}/\log n} \geqslant ca_n\sqrt{nh_{\min}^{3d}/\log n} \to \infty$$

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

Consider $\tau > 1$ case. Suppose $\zeta \leqslant d$. For any $w \in \mathcal{G}_{\vartheta}$, there exist $i \in \mathrm{N}_{\vartheta}$ and $m = 1,...,p$ such that $f_m(X_i) \geqslant a_n h_{\min}^{\zeta}$ where a_n is as defined above. For $m = 1,...,d$, set $e_m = 2h_{\min}$ if $\partial f_m(X_i)/\partial x_m \geqslant 0$ and $-2h_{\min}$ otherwise. Consider the cube $\mathcal{C}$ whose edges are parallel to axes and that contains vertices $(X_{i,1},...,X_{i,d})$ and $(X_{i,1} + 2e_1,...,X_{i,d} + 2e_d)$. By Lemma 31, for all $x \in \mathcal{C}$, $f_m(x) \geqslant ca_n h_{\min}^{\zeta}$. By the definition of N_{ϑ} and A13, there exists $l = 1,...,n$ such that $X_l \in B_{h_{\min}}(X_{i,1} + e_1,...,X_{i,d} + e_d)$. Let $s_n(w) = (l, m, h_{\min}) \in S_n$. Then $f_{s_n(w)} \geqslant ca_n h_{\min}^{\zeta}$. The rest of the proof follows from the same argument as in the case $\tau \leqslant 1$.

Suppose $\zeta > d$. The only difference between this case and the previous one is that now optimal testing bandwidth value is greater than $h_{\min}$. Let h_o be the largest bandwidth value in the set S_n that is smaller than $C(\log n/n)^{1/(2\zeta+d)}$. For any $w \in \mathcal{G}_{\vartheta}$, the same construction as above gives $s_n(w) = (l, m, h_o) \in S_n$ such that $f_m(X_j) \geqslant \rho_{\vartheta}(w, H_0) - Ch_o^{\zeta}$ for all $j = 1,...,n$ such that $X_j \in B_{h_o}(X_l)$. Since $\rho_{\vartheta}(w, H_0) \geqslant a_n(\log n/n)^{\zeta/(2\zeta+d)}$ for some sequence of real numbers a_n such that $a_n \to \infty$ as $n \to \infty$, $f_{s_n(w)} \geqslant (a_n - C)(\log n/n)^{\zeta/(2\zeta+d)}$. By Lemma 21, $V_{s_n(w)} \leqslant C/\sqrt{nh_o^d}$. Then

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

The result follows as above. □

Proof of Theorem 5. First, define functions $b_1,...,b_K$ on $(0,1]$ for $K = \lceil \tau \rceil$ by the following induction. Set $b_1(x) = +1$ for $x \in (0,1/2]$ and -1 for $x \in (1/2,1]$. Given $b_1,...,b_{k-1}$, for $i = 1,3,...,2^k - 1$ and $x \in ((i-1)2^{-k}, i2^{-k}]$, set $b_k(x) = +1$ if $b_{k-1}(y) = +1$ for $y \in ((i-1)2^{-k}, (i+1)2^{-k}]$ and -1 otherwise. For $i = 2,4,...,2^k$ and $x \in$

$((i-1)2^{-k}, i2^{-k}]$, set $b_k(x) = -1$ if $b_{k-1}(y) = +1$ for $y \in ((i-2)2^{-k}, i2^{-k}]$ and $+1$ otherwise.

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

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

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

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

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

As in the proof of Lemma 6.2 in [42], for any sequence $\phi_n = \phi_n(Y_1, ..., Y_n)$ of tests with $\sup_{w \in \mathcal{G}_0 \cap \mathcal{G}_X} E_w[\phi_n] \leqslant \alpha$,

$$\inf_{w \in \mathcal{G}_X, \rho_\vartheta(w, H_0) \geqslant Cr_n} E_w[\phi_n] - \alpha \leqslant \min_{l=1,...,N_n} E_{w_l}[\phi_n] - E_{w_0}[\phi_n] \leqslant \sum_{i=1}^{N_n} E_{w_l}[\phi_n]/N_n - E_{w_0}[\phi_n]$$
$$\leqslant E_{w_0}\left[\left(\sum_{i=1}^{N_n}(dP_{w_l}/dP_{w_0})/N_n - 1\right)\phi_n\right] \leqslant E_{w_0}\left[\left|\sum_{i=1}^{N_n}(dP_{w_l}/dP_{w_0})/N_n - 1\right|\right]$$

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

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

Note that $\omega_l \leqslant Cn^{1/2}h_n^{\tau+d/2}$. In addition, under the model w_0, ξ_l are independent standard Gaussian random variables. So, an application of Lemma 20 gives

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

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

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

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

21 and 23 gives

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

Since $a_n\sqrt{n_a h_{\max}^d \log n} \leqslant Cn^{-c}$, the bias is asymptotically negligible in comparison with the concentration rate of the test statistic. Therefore, the argument like that used in the proof of Theorem 8 leads to

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

for $P = PIA$ or RMS.

The second result follows from the same argument as in the proof of Theorem 8 since $-Ca_n \leqslant f_s \leqslant Ca_n$ ensures that the bias is again asymptotically negligible in the comparison with the concentration rate of the test statistic.

Finally, consider the third part of the corollary. If $\rho_z(w, H_0) > \rho$, then for sufficiently large n, there exists a triple $s_n = (i_n, m, h_n) \in S_n$ with h_n bounded away from zero such that $f_m(X_j) \geqslant \rho/2$ for all $X_j \in B_{h_n}(X_{i_n})$ and $\|X_{i_n}^2 - x_0\| \leqslant a_n$. The rest of the proof follows from the argument similar to that used in the proof of Theorem 9. □

Chapter 3

Central Limit Theorems and Multiplier Bootstrap when p is much larger than n

3.1 Introduction

Let $x_1, \dots, x_n$ be independent random vectors in $\mathbb{R}^p$, with each x_i having coordinates denoted by x_{ij}, i.e., $x_i = (x_{i1}, \dots, x_{ip})'$. Suppose that each x_i is centered, namely $\mathrm{E}[x_i] = 0$, and has a finite covariance matrix $\mathrm{E}[x_i x_i']$. Consider the rescaled average:

$$X := (X_1, \dots, X_p)' := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i. \tag{3.1}$$

Our goal is to obtain a distributional approximation for the statistic T_0 defined as the maximum coordinate of vector X:

$$T_0 := \max_{1 \leqslant j \leqslant p} X_j,$$

The distribution of T_0 is of interest in many applications. When p is fixed, this distribution can be approximated by the classical Central Limit Theorem (CLT) applied to X. However, in modern applications, cf. [24], p is often comparable or even larger than n, and the classical CLT does not apply in such cases. This paper provides a tractable approximation to the distribution of T_0 when p is large and possibly much larger than n.

The *first* main result of the paper is the Gaussian approximation theorem, which bounds the Kolmogorov distance between the distributions of T_0 and its Gaussian analog Z_0. Specifically, let $y_1, \dots, y_n$ be independent centered Gaussian random vectors in $\mathbb{R}^p$ such that each y_i has the same covariance matrix as x_i, namely $y_i \sim N(0, \mathrm{E}[x_i x_i'])$. Consider the rescaled average of these vectors,

$$Y := (Y_1, \dots, Y_p)' := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i. \tag{3.2}$$

Vector Y is the Gaussian analog of X in the sense of sharing the same mean and covariance matrix, namely $\mathrm{E}[X] = \mathrm{E}[Y] = 0$ and $\mathrm{E}[XX'] = \mathrm{E}[YY'] = n^{-1} \sum_{i=1}^{n} \mathrm{E}[x_i x_i']$. We then define the Gaussian analog Z_0 of T_0 as the maximum coordinate of vector Y:

$$Z_0 := \max_{1 \leqslant j \leqslant p} Y_j. \tag{3.3}$$

Our main result shows that, under suitable moment assumptions, as $n \to \infty$ and possibly $p = p_n \to \infty$,

$$\rho := \sup_{t \in \mathbb{R}} |\mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t)| \leqslant C n^{-c} \to 0, \tag{3.4}$$

where constants $c > 0$ and $C > 0$ are independent of n.

Importantly, in (3.4), p can be large in comparison to n and be nearly as large as $e^{o(n^{1/7})}$. For example, if x_{ij} are uniformly bounded (namely, $|x_{ij}| \leqslant C_1$ for some constant $C_1 > 0$ for all i and j) the Kolmogorov distance ρ converges to zero at a polynomial rate whenever $(\log p)^7/n \to 0$ at a polynomial rate. We obtain similar results when x_{ij} are sub-exponential and even non-sub-exponential under suitable moment assumptions. Figure 3.1 illustrates the result (3.4) in a non-subexponential example, which is motivated by the analysis of the Dantzig Selector of [27] in non-Gaussian settings (see Section 3.4).

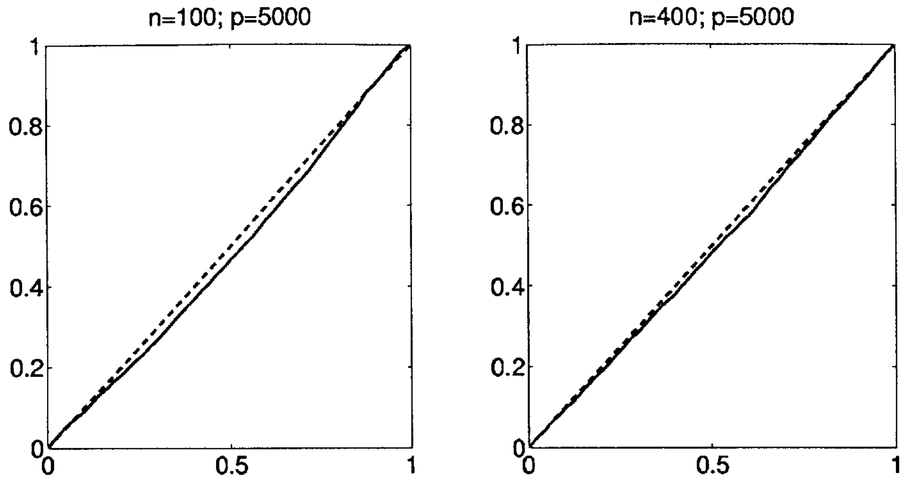


Figure 3-1: P-P plots comparing distributions of T_0 and Z_0 in the example motivated by the problem of seleting the penalty level of the Dantzig selector. Here x_{ij} are generated as $x_{ij} = z_{ij}\varepsilon_i$ with $\varepsilon_i \sim t(4)$, (a t-distribution with four degrees of freedom), and z_{ij} are non-stochastic (simulated once using $U[0,1]$ distribution independently across i and j). The dashed line is 45°. The distributions of T_0 and Z_0 are close, as (qualitatively) predicted by the CLT derived in the paper: see Corollaries 3 or 4. The quality of the Gaussian approximation is particularly good for the tail probabilities, which is most relevant for practical applications.

The proof of the Gaussian approximation result (3.4) builds on a number of technical tools such as Slepian's smart path interpolation (which is related to the solution of Stein's partial differential equation; cf. Appendix E), Stein's leave-one-out method, approximation of maxima by the smooth functions (related to "free energy" in spin glasses), and exponential inequalities for self-normalized sums. See, e.g., [106, 107, 41, 31, 108, 28, 39, 92] for introduction and discussion of some of these tools. It also critically relies on the anti-concentration and comparison bounds of maxima of Gaussian vectors derived in [34] and restated in this paper as Lemmas 32 and 34.

Our new Gaussian approximation theorem has the following innovative features. To the best of our knowledge, this is the first general result that establishes that maxima of sums of random vectors can be approximated in distribution by the maxima of sums of Gaussian random vectors when $p \gg n$ and especially when p is of order e^{n^c} for some $c > 0$. The existing techniques can also lead to results of the form (3.4) when $p = p_n \to \infty$, but under much stronger conditions on p. For example, Yurinskii's coupling implies (3.4) but requires $p^5/n \to 0$; see Example 17 (Section 10) in [94]. Second, our Gaussian approximation theorem covers cases where T_0 does not have a limit distribution as $n \to \infty$ and $p = p_n \to \infty$. In some cases, after a suitable normalization, T_0 could have an extreme value distribution as a limit distribution, but the approximation to an extreme value distribution requires some restrictions on the dependency structure among the coordinates in x_i. Our result does not require such restrictions on the dependency structure. Third, the quality of approximation in (3.4) is of polynomial order in n, which is better than the logarithmic in n quality that we could obtain in some (though not all) applications using the approximation of the distribution of T_0 by an extreme value distribution (see [71]).

Our result also contributes to the literature on multivariate central limit theorems, which are concerned with conditions under which

$$|\mathrm{P}(X \in A) - \mathrm{P}(Y \in A)| \to 0, \tag{3.5}$$

uniformly in a collection of sets A, typically *all* convex sets. Such results were developed among others, by [89, 96, 53, 17, 30], under conditions of type $p^c/n \to 0$ (also see [29]). These results rely on the anti-concentration results for Gaussian random vectors on the δ-expansions of boundaries of arbitrary convex sets A (see [14]). Note that our result also establishes (3.5), but uniformly for all convex sets of the form $A_{\max} = \{a \in \mathbb{R}^p : \max_{1 \leqslant j \leqslant p} a_j \leqslant t\}$ for $t \in \mathbb{R}$. These sets have a rather special structure that allows us to deal with $p \gg n$: in particular, concentration of measure on the δ-expansion of boundary of $A_{\max}$ is at most of order $\delta\sqrt{\log p}$ for Gaussian random vectors with unit variance, as shown in [34] (see also Lemma 32). (The relation (3.5) with $A = A_{\max}$ explains the sense in which we have a CLT, as appearing in the title of the paper.)

Note that the result (3.4) is immediately useful for inference with statistic T_0, even though $\mathrm{P}(Z_0 \leqslant t)$ needs not converge itself to a well behaved distribution function. Indeed, if the covariance matrix $n^{-1}\sum_{i=1}^n \mathrm{E}[x_i x_i']$ is known, then $c_{Z_0}(1-\alpha) := (1-\alpha)$-

quantile of Z_0, can be computed numerically, and we have

$$|\mathrm{P}(T_0 \leqslant c_{Z_0}(1-\alpha)) - (1-\alpha)| \leqslant Cn^{-c} \to 0. \tag{3.6}$$

A chief application of this kind arises in determination of the penalty level for the Dantzig selector of [27] in the high-dimensional regression with non-Gaussian errors, which we examine in Section 5. There, under the canonical (homoscedastic) noise, the covariance matrix is known, and so quantiles of Z_0 can be easily computed numerically and used for choosing the penalty level. However, if the noise is heteroscedastic, the covariance matrix is no longer known, and this approach is no longer feasible. This motivates our second main result.

The *second* main result of the paper establishes validity of the multiplier bootstrap for estimating quantiles of Z_0 when the covariance matrix $n^{-1}\sum_{i=1}^n \mathrm{E}[x_i x_i']$ is unknown. More precisely, we define the Gaussian-symmetrized version W_0 of T_0 by multiplying x_i with i.i.d. standard Gaussian random variables $e_1, \dots, e_n$:

$$W_0 := \max_{1\leqslant j\leqslant p} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij} e_i. \tag{3.7}$$

We show that the conditional quantiles of W_0 given data $(x_i)_{i=1}^n$ are able to consistently estimate the quantiles of Z_0 and hence those of T_0 (where the notion of consistency used is the one that guarantees asymptotically valid inference). Here the primary factor driving the bootstrap estimation error is the maximum difference between the empirical and population covariance matrices:

$$\Delta := \max_{1\leqslant j,k\leqslant p} \left| \frac{1}{n} \sum_{i=1}^n (x_{ij}x_{ik} - \mathrm{E}[x_{ij}x_{ik}]) \right|,$$

which can converge to zero even when p is much larger than n. For example, when x_{ij} are uniformly bounded, the multiplier bootstrap is valid for inference if $(\log p)^7/n \to 0$. Earlier related results on bootstrap in the "$p \to \infty$ but $p/n \to 0$" regime were obtained in [78]; interesting results for the case $p \gg n$ based on concentration inequalities and symmetrization are studied in [8, 9], albeit the approach and results are quite different from those given here. In particular, in [8], either Gaussianity or symmetry in distribution is imposed on the data.

The key motivating example of our analysis is the high-dimensional sparse regression model. In this model, [27] and [19] assume Gaussian errors to analyze the Dantzig selector and Lasso. Our results show that Gaussianity is not necessary and the Gaussian-like conclusions hold approximately, with just the fourth moment of the regression errors being bounded. Moreover, our approximation allows to take into account correlations among the regressors. This leads to a better choice of the penalty level and tighter bounds on performance than those that had been available previously. For example, some of the same goals had been accomplished using moderate deviations for self-normalized sums, combined with the union bound [16]. However, the union bound does not take into account correlations among the regressors, and so it may be overly conservative in some applications.

Our results have a broad range of other applications. In addition to the high-dimensional estimation example, we show in the Supplemental Material how to apply our results in the multiple hypothesis testing via the step-down method of [103] and to specification testing. In either case number of hypotheses to be tested or the number of moment restrictions to be tested can be much larger than the sample size. Lastly, in a companion work ([32]), we are exploring the strong coupling for suprema of general empirical processes based on the methods developed here and maximal inequalities. These results represent a useful complement to the results based on the Hungarian coupling developed by [69, 22, 67, 98] for the entire empirical process and have applications to inference in nonparametric problems such as construction of uniform confidence bands (see, e.g., [51]).

3.1.1 Organization of the paper

In Section 3.2, we give the results on Gaussian approximation, and in Section 3.3 on the multiplier bootstrap. In Section 3.4, we present an application to the Dantzig selector. Appendices 3.5-3.8 contain proofs for each of these sections, with Appendix

3.5 stating auxiliary tools and lemmas. Due to the space limitation, we put additional results and proofs into Supplemental Material, Appendices 3.11-3.10. In particular, Appendices 3.11 and 3.12 provide additional applications to multiple hypothesis and adaptive specification testing.

3.1.2 Notation

In what follows, unless otherwise stated, we will assume that $p \geqslant 3$. In making asymptotic statements we assume that $n \to \infty$ with understanding that p depends on n and possibly $p \to \infty$ as $n \to \infty$. Constants $c, C, c_1, C_1, c_2, C_2, \dots$ are understood to be independent of n. Throughout the paper, $\mathbb{E}_n[\cdot]$ denotes the average over index $1 \leqslant i \leqslant n$, i.e., it simply abbreviates the notation $n^{-1}\sum_{i=1}^n[\cdot]$. E.g., $\mathbb{E}_n[x_{ij}^2] = n^{-1}\sum_{i=1}^n x_{ij}^2$. In addition, $\bar{\mathrm{E}}[\cdot] = \mathbb{E}_n[\mathrm{E}[\cdot]]$. For example, $\bar{\mathrm{E}}[x_{ij}^2] = n^{-1}\sum_{i=1}^n \mathrm{E}[x_{ij}^2]$. For a function $f : \mathbb{R} \to \mathbb{R}$, we write $\partial^k f(x) = \partial^k f(x)/\partial x^k$ for nonnegative integer k; for a function $f : \mathbb{R}^p \to \mathbb{R}$, we write $\partial_j f(x) = \partial f(x)/\partial x_j$ for $j = 1, \dots, p$, where $x = (x_1, \dots, x_p)'$. Denote by $C^k(\mathbb{R})$ the class of k times continuously differentiable functions from $\mathbb{R}$ to itself, and denote by $C_b^k(\mathbb{R})$ the class of all functions $f \in C^k(\mathbb{R})$ such that $\sup_{z \in \mathbb{R}} |\partial^j f(z)| < \infty$ for $j = 0, \dots, k$. We write $a \lesssim b$ if a is smaller than or equal to b up to a universal positive constant. For $a, b \in \mathbb{R}$, we write $a \vee b = \max\{a, b\}$.

3.2 Central Limit Theorems for Maxima of Non-Gaussian Sums

3.2.1 Comparison Theorems and Non-Asymptotic Gaussian Approximations

The purpose of this section is to compare and bound the difference between the expectations and distribution functions of the non-Gaussian to Gaussian maxima:

$$T_0 := \max_{1 \leqslant j \leqslant p} X_j \quad \text{and} \quad Z_0 := \max_{1 \leqslant j \leqslant p} Y_j,$$

where vector X is defined in equation (3.1) and Y in equation (3.2). Here and in what follows, without loss of generality, we will assume that $(x_i)_{i=1}^n$ and $(y_i)_{i=1}^n$ are independent. The following envelopes and bounds on moments will be used in stating the bounds in Gaussian approximations:

$$S_i := \max_{1\leqslant j\leqslant p}(|x_{ij}| + |y_{ij}|), \quad M_k := \max_{1\leqslant j\leqslant p}(\bar{\mathrm{E}}[x_{ij}^k])^{1/k}. \tag{3.8}$$

The problem of comparing distributions of maxima is of intrinsic difficulty since the maximum function $z = (z_1, \dots, z_p)' \mapsto \max_{1\leqslant j\leqslant p} z_j$ is non-differentiable. To circumvent the problem, we use a smooth approximation of the maximum function. For $z = (z_1, \dots, z_p)' \in \mathbb{R}^p$, consider the function:

$$F_\beta(z) := \beta^{-1} \log\left(\sum_{j=1}^p \exp(\beta z_j)\right),$$

which approximates the maximum function, where $\beta > 0$ is the smoothing parameter that controls the level of approximation (we call this function the "smooth max function"). Indeed, an elementary calculation shows that for all $z \in \mathbb{R}^p$,

$$0 \leqslant F_\beta(z) - \max_{1\leqslant j\leqslant p} z_j \leqslant \beta^{-1}\log p. \tag{3.9}$$

This smooth max function arises in the definition of "free energy" in spin glasses; see, e.g., [108].

We start with the following "warm-up" theorem that conveys the main qualitative feature of the problem. Here and in what follows, for a smooth function $g : \mathbb{R} \to \mathbb{R}$, write

$$G_k := \sup_{z\in\mathbb{R}} |\partial^k g(z)|, \ k \geqslant 0.$$

Theorem 13 (Comparison of Gaussian to Non-Gaussian Maxima, I). *For every $g \in C_b^3(\mathbb{R})$ and $\beta > 0$,*

$$|\mathrm{E}[g(F_\beta(X)) - g(F_\beta(Y))]| \lesssim n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)\bar{\mathrm{E}}[S_i^3],$$

and hence

$$|\mathrm{E}[g(T_0) - g(Z_0)]| \lesssim n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)\bar{\mathrm{E}}[S_i^3] + \beta^{-1}G_1 \log p.$$

Comment 13 (Optimizing the bound). *The theorem bounds the difference between the expectations of smooth functions of maxima. The optimal value of the last bound is given by*

$$\min_{\beta>0} n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)\bar{\mathrm{E}}[S_i^3] + \beta^{-1}G_1 \log p.$$

We postpone choices of β to the proofs of subsequent corollaries, leaving ourselves more flexibility in optimizing bounds in those corollaries. □

Deriving a bound on the Kolmogorov distance between distributions of T_0 and Z_0 from Theorem 13 is *not* a trivial issue and this step relies on the following *anti-concentration* inequality for maxima of Gaussian random variables, which is derived in [34].

Lemma 32 (Anti-Concentration). *Let $\xi_1, \dots, \xi_p$ be (not necessarily independent) centered Gaussian random variables with $\sigma_j^2 := \mathrm{E}[\xi_j^2] > 0$ for all $1 \leqslant j \leqslant p$. Let $\underline{\sigma} = \min_{1\leqslant j\leqslant p} \sigma_j$ and $\bar{\sigma} = \max_{1\leqslant j\leqslant p} \sigma_j$. Then for every $\varsigma > 0$,*

$$\sup_{z\in\mathbb{R}} \mathrm{P}\left(|\max_{1\leqslant j\leqslant p} \xi_j - z| \leqslant \varsigma \right) \leqslant C\varsigma\sqrt{1 \vee \log(p/\varsigma)},$$

where $C > 0$ is a constant depending only on $\underline{\sigma}$ and $\bar{\sigma}$. When σ_j are all equal, $\log(p/\varsigma)$ *on the right side can be replaced by* $\log p$.

By Theorem 13 and Lemma 32, we can now derive a bound on the Kolmogorov distance between distributions of T_0 and Z_0.

Corollary 3 (Central Limit Theorem, I). *Suppose that there are some constants $c_1 > 0$ and $C_1 > 0$ such that $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for all $1 \leqslant j \leqslant p$. Then there exists a constant $C > 0$ depending only on c_1 and C_1 such that*

$$\rho := \sup_{t\in\mathbb{R}} |\mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t)| \leqslant C(n^{-1}(\log(pn))^7)^{1/8}(\bar{\mathrm{E}}[S_i^3])^{1/4}.$$

Comment 14 (Main qualitative feature: logarithmic dependence on p). *Theorem 13 and Corollary 3 imply that the error of approximating the maximum coordinate in the sum of independent random vectors by its Gaussian analogue depends on p (possibly) only through* $\log p$. *This is the main qualitative feature of all the results in this paper. Note also that the term* $\bar{\mathrm{E}}[S_i^3]$ *implicitly encodes the complexity of the vectors, in particular it will reflect the correlation structure of vectors X and Y. However, both Theorem 13 and Corollary 3 and all subsequent results given below do not limit the dependence among the coordinates in x_i.* □

Comment 15 (Motivation for the next result). *While Theorem 13 and Corollary 3 convey an important qualitative aspect of the problem and admit easy-to-grasp proofs, an important disadvantage of these results is that the bounds depend on* $\bar{\mathrm{E}}[S_i^3]$. *If* $\bar{\mathrm{E}}[S_i^3] \leqslant C$, *Corollary 3 leads to* $\rho = O((n^{-1}(\log(pn))^7)^{1/8})$ *and* $\rho \to 0$ *as long as* $\log p = o(n^{1/7})$. *This is the case when, for example, as in caption to Figure 1,*

$$x_{ij} = z_{ij}\varepsilon_i, \quad z_{ij} \text{ are non-stochastic with } |z_{ij}| \leqslant C, \quad \mathrm{E}[|\varepsilon_i|^3] \leqslant C.$$

When $\bar{\mathrm{E}}[S_i^3]$ *increases with n, however, the bounds need not be as good, and can be improved considerably by using a truncation method. Using such a method in conjunction with the proof strategy of Theorem 13, we derive in Theorem 14 below a bound that can be much better in the latter scenario. The improvement here comes at a cost of a more involved statement, involving truncation parameters.* □

To derive our next main result, we employ a truncation method. Given a threshold level $u > 0$, define a truncated version of x_{ij} by

$$\tilde{x}_{ij} = x_{ij} 1\left\{|x_{ij}| \leqslant u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\right\} - \mathrm{E}\left[x_{ij} 1\left\{|x_{ij}| \leqslant u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\right\}\right]. \tag{3.10}$$

Let $\varphi_x(u)$ be the infimum, which is attained, over all numbers $\varphi \geqslant 0$ such that

$$\bar{\mathrm{E}}\left[x_{ij}^2 1\left\{|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\right\}\right] \leqslant \varphi^2 \bar{\mathrm{E}}[x_{ij}^2]. \tag{3.11}$$

Note that the function $\varphi_x(u)$ is right-continuous; it measures the impact of truncation on second moments. Define $u_x(\gamma)$ as the infimum $u \geqslant 0$ such that

$$\mathrm{P}\left(|x_{ij}| \leqslant u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p\right) \geqslant 1-\gamma.$$

Also define $\varphi_y(u)$ and $u_y(\gamma)$ by the corresponding quantities for the analogue Gaussian case, namely with $(x_i)_{i=1}^n$ replaced by $(y_i)_{i=1}^n$ in the above definitions. Throughout the paper we use the following quantities:

$$\varphi(u) := \varphi_x(u) \vee \varphi_y(u), \quad u(\gamma) := u_x(\gamma) \vee u_y(\gamma).$$

Here is the main theorem of this section. Recall the definition of M_k in (3.8).

Theorem 14 (Comparison of Gaussian to Non-Gaussian Maxima, II). *Let $\beta > 0, u > 0$ and $\gamma \in (0,1)$ be such that $2\sqrt{2}uM_2\beta/\sqrt{n} \leqslant 1$ and $u \geqslant u(\gamma)$. Then for every $g \in C_b^3(\mathbb{R})$,*

$$|\mathrm{E}[g(F_\beta(X)) - g(F_\beta(Y))]| \lesssim D_n(g,\beta,u,\gamma),$$

and hence

$$|\mathrm{E}[g(T_0) - g(Z_0)]| \lesssim D_n(g,\beta,u,\gamma) + \beta^{-1}G_1\log p,$$

where

$$\begin{aligned} D_n(g,\beta,u,\gamma) := n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)M_3^3 + (G_2 + \beta G_1)M_2^2\varphi(u) \\ + G_1M_2\varphi(u)\sqrt{\log(p/\gamma)} + G_0\gamma. \end{aligned}$$

By Theorem 14 and Lemma 32, we can obtain a bound on the Kolmogorov distance between the distribution functions of T_0 and Z_0.

Corollary 4 (Central Limit Theorem, II). *Suppose that there are some constants*

$0 < c_1 < C_1$ *such that* $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ *for* $1 \leqslant j \leqslant p$. *Then for every* $\gamma \in (0,1)$,

$$\rho \leqslant C\left[n^{-1/8}(M_3^{3/4} \vee M_4^{1/2})(\log(pn/\gamma))^{7/8} + n^{-1/2}(\log(pn/\gamma))^{3/2}u(\gamma) + \gamma\right],$$

where $C > 0$ *is a constant that depends on* c_1 *and* C_1 *only.*

In applications it is useful to bound the upper function $u(\gamma)$. Here is a simple and effective way of doing this. Let $h : [0,\infty) \to [0,\infty)$ be a *Young-Orlicz modulus*, i.e., a convex and strictly increasing function with $h(0) = 0$. Denote by h^{-1} the inverse function of h. Standard examples include the power function $h(v) = v^q$ with inverse $h^{-1}(\gamma) = \gamma^{1/q}$ and the exponential function $h(v) = \exp(v) - 1$ with inverse $h^{-1}(\gamma) = \log(\gamma + 1)$. These functions describe how many moments the random variables have, for example, a random variable ξ has finite q-th moment if $\mathrm{E}[|\xi|^q] < \infty$, and is sub-exponential if $\mathrm{E}[\exp(|\xi|/C)] < \infty$ for some $C > 0$. We refer to [111], Chapter 2.2, for further details on Young-Orlicz moduli.

Lemma 33 (Bounds on the upper function $u(\gamma)$). *Let* $h : [0,\infty) \to [0,\infty)$ *be a Young-Orlicz modulus, and let* $B > 0$ *and* $D > 0$ *be constants such that* $(\mathrm{E}[x_{ij}^2])^{1/2} \leqslant B$ *for all* $1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p$, *and* $\bar{\mathrm{E}}[h(\max_{1\leqslant j\leqslant p}|x_{ij}|/D)] \leqslant 1$. *Then under the condition of Corollary 4,*

$$u(\gamma) \leqslant C\max\{Dh^{-1}(n/\gamma), B\sqrt{\log(pn/\gamma)}\},$$

where $C > 0$ *is a constant that depends on* c_1 *and* C_1 *only.*

In applications, parameters B and D (with M_3 and M_4 as well) are allowed to increase with n. The size of these parameters and the choice of the Young-Orlicz modulus are case-specific.

3.2.2 Examples of Applications

The purpose of this subsection is to obtain bounds on ρ for various leading examples frequently encountered in applications. We are concerned with simple conditions under which ρ decays polynomially in n.

Let $c_1 > 0, c_2 > 0, C_1 > 0$ be some constants, and let $B_n \geqslant 1$ be a sequence of constants. We allow for the case where $B_n \to \infty$ as $n \to \infty$. We shall first consider applications where one of the following conditions is satisfied *uniformly in* $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$:

(E.1) $\bar{\mathrm{E}}[x_{ij}^2] \geqslant c_1$ and $\bar{\mathrm{E}}[S_i^3] \leqslant C_1$;

(E.2) $\bar{\mathrm{E}}[x_{ij}^2] \geqslant c_1$ and $\mathrm{E}[\exp(|x_{ij}|/C_1)] \leqslant 2$;

(E.3) $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ and $|x_{ij}| \leqslant B_n$.

Comment 16. *Condition (E.1) is perhaps the simplest example in this paper; under this condition application of Corollary 3 is effective. A concrete example with condition (E.1) satisfied is the case where $x_{ij} = z_{ij}\varepsilon_i$, z_{ij} are non-stochastic with $|z_{ij}| \leqslant C$, and $\mathrm{E}[|\varepsilon_i|^3] \leqslant C$. Conditions (E.2)-(E.5) are more elaborate, intended to cover cases where moments of the envelopes S_i and higher order moments M_3 and M_4 increase with n. In these cases the use of Corollary 3 is not effective, and we shall use Corollary 4 instead. Condition (E.2) covers vectors x_i made up from sub-exponential random variables, including sub-Gaussian as a special case; this example is quite often used in high-dimensional statistics. Condition (E.3) covers variables that are bounded by B_n, which may increase with n; many applications, after a suitable truncation, can be covered by it.* □

We shall also consider regression applications where one of the following conditions is satisfied *uniformly in* $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$:

(E.4) $x_{ij} = z_{ij}\varepsilon_{ij}$, where z_{ij} are non-stochastic with $|z_{ij}| \leqslant B_n$, $\mathbb{E}_n[z_{ij}^2] = 1$, and $\mathrm{E}[\varepsilon_{ij}] = 0$, $\mathrm{E}[\varepsilon_{ij}^2] \geqslant c_1$, and $\mathrm{E}[\exp(|\varepsilon_{ij}|/C_1)] \leqslant 2$; or

(E.5) $x_{ij} = z_{ij}\varepsilon_{ij}$, where z_{ij} are non-stochastic with $|z_{ij}| \leqslant B_n$, $\mathbb{E}_n[z_{ij}^2] = 1$, and $\mathrm{E}[\varepsilon_{ij}] = 0$, $\mathrm{E}[\varepsilon_{ij}^2] \geqslant c_1$, and $\mathrm{E}[\max_{1 \leqslant j \leqslant p} \varepsilon_{ij}^4] \leqslant C_1$.

Comment 17. *The last two cases cover examples that arise in high-dimensional regression, e.g., [27], which we shall revisit later in the paper. Typically, ε_{ij} are independent of j (i.e., $\varepsilon_{ij} = \varepsilon_i$) and hence* $\mathrm{E}[\max_{1 \leqslant j \leqslant p} \varepsilon_{ij}^4] \leqslant C_1$ *in condition (E.5) reduces* *to* $\mathrm{E}[\varepsilon_i^4] \leqslant C_1$ *(we allow* ε_{ij} *dependent on* j *so that Corollary 5 covers the multiple hypothesis testing example in Appendix 3.11). Interestingly, these examples are also connected to spin glasses, see e.g., [108] and [92]* (z_{ij} *can be interpreted as generalized products of "spins" and* ε_i *as their random "interactions").* □

Corollary 5 **(Central Limit Theorem in Leading Examples).** *Suppose that one of the following conditions is satisfied: (i) condition (E.1) and* $(\log(pn))^7/n \leqslant C_1 n^{-c_2}$*; (ii) condition (E.2) and* $(\log(pn))^7/n \leqslant C_1 n^{-c_2}$*; (iii) condition (E.3) and* $B_n^2(\log(pn))^7/n \leqslant C_1 n^{-c_2}$*; (vi) condition (E.4) and* $B_n^2(\log(pn))^7/n \leqslant C_1 n^{-c_2}$*; or (v) condition (E.5) and* $B_n^4(\log(pn))^7/n \leqslant C_1 n^{-c_2}$*. Then there exist constants* $c > 0$ *and* $C > 0$ *depending only on* c_1, c_2 *and* C_1 *such that*

$$\rho \leqslant Cn^{-c}.$$

Comment 18. *Cases (ii)-(v) indeed follow relatively directly from Corollary 4 with help of Lemma 33. Moreover, from Lemma 33, it is routine to find other conditions that lead to the conclusion of Corollary 5.* □

3.3 Multiplier Bootstrap

3.3.1 A Gaussian-to-Gaussian Comparison Theorem

The proofs of the main results in this section rely on the following lemma. Let V and Y be centered Gaussian random vectors in $\mathbb{R}^p$ with covariance matrices Σ^V and Σ^Y, respectively. The following lemma compares the distribution functions of $\max_{1\leqslant j\leqslant p} V_jand\max_{1\leqslant j\leqslant p} Y_j$ in terms of p and

$$\Delta_0 := \max_{1\leqslant j,k\leqslant p} \left|\Sigma_{jk}^V - \Sigma_{jk}^Y\right|.$$

Lemma 34 (Comparison of Distributions of Gaussian Maxima). *Suppose that there are some constants* $0 < c_1 < C_1$ *such that* $c_1 \leqslant \Sigma_{jj}^Y \leqslant C_1$ *for all* $1 \leqslant j \leqslant p$*. Then* *there exists a constant $C > 0$ depending only on c_1 and C_1 such that*

$$\sup_{t\in\mathbb{R}}\left|\mathrm{P}\left(\max_{1\leqslant j\leqslant p} V_j \leqslant t\right) - \mathrm{P}\left(\max_{1\leqslant j\leqslant p} Y_j \leqslant t\right)\right| \leqslant C\Delta_0^{1/3}(1 \vee \log(p/\Delta_0))^{2/3}.$$

Comment 19. *The result is derived in [34], and extends that of [28] who gave an explicit error in Sudakov-Fernique comparison of expecations of maxima of Gaussian vectors.* □

3.3.2 Multiplier Bootstrap Theorems

Suppose that we have a dataset $(x_i)_{i=1}^n$ consisting of n independent centered random vectors x_i in $\mathbb{R}^p$. In this section we are interested in approximating quantiles of

$$T_0 = \max_{1\leqslant j\leqslant p}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{ij} \tag{3.12}$$

using the multiplier bootstrap method. Specifically, let $(e_i)_{i=1}^n$ be a sequence of i.i.d. $N(0,1)$ variables independent of $(x_i)_{i=1}^n$, and let

$$W_0 = \max_{1\leqslant j\leqslant p}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{ij}e_i. \tag{3.13}$$

Then we define the multiplier bootstrap estimator of the α-quantile of T_0 as the conditional α-quantile of W_0 given $(x_i)_{i=1}^n$, i.e.,

$$c_{W_0}(\alpha) := \inf\{t \in \mathbb{R} : \mathrm{P}_e(W_0 \leqslant t) \geqslant \alpha\},$$

where P_e is the probability measure induced by the multiplier variables $(e_i)_{i=1}^n$ holding $(x_i)_{i=1}^n$ fixed (i.e., $\mathrm{P}_e(W_0 \leqslant t) = \mathrm{P}(W_0 \leqslant t \mid (x_i)_{i=1}^n)$). The multiplier bootstrap theorem below provides a non-asymptotic bound on the bootstrap estimation error:

$$\left|\mathrm{P}(T_0 \leqslant c_{W_0}(\alpha)) - \alpha\right|.$$

Before presenting the theorem, we first give a simple useful lemma that is helpful in the proof of the theorem and in power analysis in applications. Define

$$c_{Z_0}(\alpha) := \inf\{t \in \mathbb{R} : \mathrm{P}(Z_0 \leqslant t) \geqslant \alpha\},$$

where $Z_0 = \max_{1\leqslant j\leqslant p} \sum_{i=1}^n y_{ij}/\sqrt{n}$ and $(y_i)_{i=1}^n$ is a sequence of independent $N(0, \mathrm{E}[x_i x_i'])$ vectors. Recall that

$$\Delta = \max_{1\leqslant j,k\leqslant p} \left|\mathbb{E}_n[x_{ij}x_{ik}] - \bar{\mathrm{E}}[x_{ij}x_{ik}]\right|.$$

Lemma 35 (Comparison of Quantiles, I). *Suppose that there are some constants* $0 < c_1 < C_1$ *such that* $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ *for all* $1 \leqslant j \leqslant p$. *Then for every* $\alpha \in (0,1)$,

$$\mathrm{P}\big(c_{W_0}(\alpha) \leqslant c_{Z_0}(\alpha + \pi(\vartheta))\big) \geqslant 1 - \mathrm{P}(\Delta > \vartheta),$$
$$\mathrm{P}\big(c_{Z_0}(\alpha) \leqslant c_{W_0}(\alpha + \pi(\vartheta))\big) \geqslant 1 - \mathrm{P}(\Delta > \vartheta),$$

where, for $C_2 > 0$ *denoting a constant depending only on* c_1 *and* C_1,

$$\pi(\vartheta) := C_2 \vartheta^{1/3}(1 \vee \log(p/\vartheta))^{2/3}.$$

Recall that $\rho := \sup_{t\in\mathbb{R}} |\mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t)|$. We are now in position to state the main theorem of this section.

Theorem 15 (Validity of Multiplier Bootstrap, I). *Suppose that for some constants* $0 < c_1 < C_1$, *we have* $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ *for all* $1 \leqslant j \leqslant p$. *Then for any* $\vartheta > 0$,

$$\sup_{\alpha\in(0,1)} |\mathrm{P}(T_0 \leqslant c_{W_0}(\alpha)) - \alpha| \leqslant \rho + \pi(\vartheta) + \mathrm{P}(\Delta > \vartheta).$$

Theorem 15 provides a useful result for the case where the statistics are maxima of exact averages. There are many applications, however, where the relevant statistics arise as maxima of approximate averages. The following result shows that the theorem continues to apply if the approximation error of the relevant statistic by a maximum of an exact average can be suitably controlled. Specifically, suppose that a statistic of interest, say $T = T(x_1 \ldots, x_n)$ which may not be of the form (3.12), can be approximated by T_0 of the form (3.12), and that the multiplier bootstrap is performed on a statistic $W = W(x_1, \ldots, x_n, e_1, \ldots, e_n)$, which may be different from (3.13) but still can be approximated by W_0 of the form (3.13).

We require the approximation to hold in the following sense: there exist $\zeta_1 \geqslant 0$ and $\zeta_2 \geqslant 0$, depending on n (and typically $\zeta_1 \to 0, \zeta_2 \to 0$ as $n \to \infty$), such that

$$\mathrm{P}(|T - T_0| > \zeta_1) < \zeta_2, \tag{3.14}$$

$$\mathrm{P}(\mathrm{P}_e(|W - W_0| > \zeta_1) > \zeta_2) < \zeta_2. \tag{3.15}$$

We use the α-quantile of $W = W(x_1, \ldots, x_n, e_1, \ldots, e_n)$, computed conditional on $(x_i)_{i=1}^n$:

$$c_W(\alpha) := \inf\{t \in \mathbb{R} : \mathrm{P}_e(W \leqslant t) \geqslant \alpha\},$$

as an estimate of the α-quantile of T.

Lemma 36 (Comparison of Quantiles, II). *Suppose that condition (3.15) is satisfied. Then for every $\alpha \in (0,1)$,*

$$\mathrm{P}(c_W(\alpha) \leqslant c_{W_0}(\alpha + \zeta_2) + \zeta_1) \geqslant 1 - \zeta_2,$$
$$\mathrm{P}(c_{W_0}(\alpha) \leqslant c_W(\alpha + \zeta_2) + \zeta_1) \geqslant 1 - \zeta_2.$$

The next result provides a bound on the bootstrap estimation error.

Theorem 16 (Validity of Multiplier Bootstrap, II). *Suppose that, for some constants $0 < c_1 < C_1$, we have $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for all $1 \leqslant j \leqslant p$. Moreover, suppose that conditions (3.14) and (3.15) are satisfied. Then for any $\vartheta > 0$,*

$$\sup_{\alpha \in (0,1)} |\mathrm{P}(T \leqslant c_W(\alpha)) - \alpha| \leqslant \rho + \pi(\vartheta) + \mathrm{P}(\Delta > \vartheta) + C_3 \zeta_1 \sqrt{1 \vee \log(p/\zeta_1)} + \zeta_2,$$

where $\pi(\cdot)$ is defined in Lemma 35, and $C_3 > 0$ depends only on c_1 and C_1.

3.3.3 Examples of Applications: Revisited

Here we revisit the examples in Section 3.2.2 and see how the multiplier bootstrap works for these leading examples. Let, as before, $c_1 > 0, c_2 > 0$ and $C_1 > 0$ be some constants, and let $B_n \geqslant 1$ be a sequence of constants. Recall conditions (E.2)-(E.5) in Section 3.2.2.

Corollary 6 (Multiplier Bootstrap in Leading Examples). *Suppose that conditions (3.14) and (3.15) hold with $\zeta_1\sqrt{\log p} + \zeta_2 \leqslant C_1 n^{-c_2}$. Moreover, suppose that one of the following conditions is satisfied: (i) condition (E.2) and $(\log(pn))^7/n \leqslant C_1 n^{-c_2}$; (ii) condition (E.3), and $B_n^2(\log(pn))^7/n \leqslant C_1 n^{-c_2}$; (iii) condition (E.4) and $B_n^2(\log(pn))^7/n \leqslant C_1 n^{-c_2}$; or (iv) condition (E.5) and $B_n^4(\log(pn))^7/n \leqslant C_1 n^{-c_2}$. Then there exist constants $c > 0$ and $C > 0$ depending only on c_1, c_2 and C_1 such that*

$$\sup_{\alpha \in (0,1)} |\mathrm{P}(T \leqslant c_W(\alpha)) - \alpha| \leqslant C n^{-c}.$$

Comment 20. *This corollary shows that the multiplier bootstrap is valid with a polynomial rate of accuracy for the significance level under weak conditions. This is in contrast with the extremal theory of Gaussian processes that provides only a logarithmic rate of approximation (see, e.g., [71] and [56]).* □

3.4 Application: Dantzig Selector in the Non-Gaussian Model

The purpose of this section is to demonstrate the case with which the CLT and the multiplier bootstrap theorem given in Corollaries 5 and 6 can be applied in important problems, dealing with a high-dimensional inference and estimation. We consider the Dantzig selector previously studied in the path-breaking works of [27], [19], [114] in the Gaussian setting and of [68] in a sub-exponential setting. Here we consider the non-Gaussian case, where the errors have only four bounded moments, and derive the performance bounds that are approximately as sharp as in the Gaussian model. We consider both homoscedastic and heteroscedastic models.

3.4.1 Homoscedastic case

Let $(z_i, y_i)_{i=1}^n$ be a sample of independent observations where $z_i \in \mathbb{R}^p$ is a non-stochastic vector of regressors. We consider the model

$$y_i = z_i'\beta + \varepsilon_i, \quad \mathrm{E}[\varepsilon_i] = 0, \ i = 1, \dots, n, \ \mathbb{E}_n[z_{ij}^2] = 1, \ j = 1, \dots, p,$$

where y_i is a random scalar dependent variable, and the regressors are normalized in such a way that $\mathbb{E}_n[z_{ij}^2] = 1$. Here we consider the homoscedastic case:

$$\mathrm{E}[\varepsilon_i^2] = \sigma^2, \ i = 1, \dots, n,$$

where σ^2 is assumed to be known (for simplicity). We allow p to be substantially larger than n. It is well known that a condition that gives a good performance for the Dantzig selector is that β is sparse, namely $\|\beta\|_0 \leqslant s \ll n$ (although this assumption will not be invoked below explicitly).

The aim is to estimate the vector β in some semi-norms of interest: $\|\cdot\|_I$. For example, given an estimator $\widehat{\beta}$ the prediction semi-norm for $\delta = \widehat{\beta} - \beta$ is

$$\|\delta\|_{\mathrm{pr}} = \sqrt{\mathbb{E}_n[(z_i'\delta)^2]},$$

or the j-th component seminorm for δ is

$$\|\delta\|_{\mathrm{jc}} = |\delta_j|,$$

and so on. The label I designates the name of a norm of interest.

The Dantzig selector is the estimator defined by

$$\widehat{\beta} \in \arg\min_{b \in \mathbb{R}^p} \|b\|_{\ell_1} \text{ subject to } \sqrt{n} \max_{1 \leqslant j \leqslant p} |\mathbb{E}_n[z_{ij}(y_i - z_i'b)]| \leqslant \lambda, \qquad (3.16)$$

where $\|\beta\|_{\ell_1} = \sum_{j=1}^{p} |\beta_j|$ is the ℓ_1-norm. An ideal choice of the penalty level λ is meant to ensure that

$$T_0 := \sqrt{n} \max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}\varepsilon_i]| \leqslant \lambda$$

with a prescribed probability $1-\alpha$. Hence we would like to set penalty level λ equal to

$$c_{T_0}(1-\alpha) := (1-\alpha)\text{-quantile of } T_0,$$

(note that z_i are treated as fixed). Indeed, this penalty would take into account the correlation amongst the regressors, thereby adapting the performance of the estimator to the design condition. We can approximate this quantity using the central limit theorems derived in Section 2. Specifically, let

$$Z_0 := \sigma\sqrt{n} \max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}e_i]|,$$

where e_i are i.i.d. $N(0,1)$ random variables independent of the data. We then estimate $c_{T_0}(1-\alpha)$ by

$$c_{Z_0}(1-\alpha) := \ (1-\alpha)\text{-quantile of } Z_0.$$

Note that we can calculate $c_{Z_0}(1-\alpha)$ numerically with any specified precision by the simulation. (In a Gaussian model, design-adaptive penalty level $c_{Z_0}(1-\alpha)$ was proposed in [15], but its extension to non-Gaussian cases was not available up to now).

An alternative choice of the penalty level is given by

$$c_0(1-\alpha) := \sigma\Phi^{-1}(1-\alpha/(2p)),$$

which is the canonical choice; see [27] and [19]. Note that canonical choice $c_0(1-\alpha)$ disregards the correlation amongst the regressors, and is therefore more conservative than $c_{Z_0}(1-\alpha)$. Indeed, by the union bound, we see that

$$c_{Z_0}(1-\alpha) \leqslant c_0(1-\alpha).$$

Our first result below shows that the *either* of the two penalty choices, $\lambda = c_{Z_0}(1-\alpha)$ or $\lambda = c_0(1-\alpha)$, are approximately valid under non-Gaussian noise–under the mild moment assumption $\mathrm{E}[\varepsilon_i^4] \leqslant$ const. replacing the canonical Gaussian noise assumption. To derive this result we apply our CLT to T_0 to establish that the difference between distribution functions of T_0 and Z_0 approaches zero at polynomial speed. Indeed T_0 can be represented as a maximum of averages, $T_0 = \max_{1\leqslant k\leqslant 2p} n^{-1/2}\sum_{i=1}^n \tilde{z}_{ik}\varepsilon_i$, for $\tilde{z}_i = (z_i', -z_i')'$, and therefore our CLT applies.

To derive the bound on estimation error $\|\delta\|_I$ in a seminorm of interest, we employ the following identifiability factor:

$$\kappa_I(\beta) := \inf_{\delta\in\mathbb{R}^p}\left\{\max_{1\leqslant j\leqslant p}\frac{|\mathbb{E}_n[z_{ij}(z_i'\delta)]|}{\|\delta\|_I} : \delta \in \mathcal{R}(\beta), \|\delta\|_I \neq 0\right\},$$

where $\mathcal{R}(\beta) := \{\delta \in \mathbb{R}^p : \|\beta+\delta\|_{\ell_1} \leqslant \|\beta\|_{\ell_1}\}$ is the restricted set; $\kappa_I(\beta)$ is defined as ∞ if $\mathcal{R}(\beta) = \{0\}$ (this happens if $\beta = 0$). The factors summarize the impact of sparsity of true parameter value β and the design on the identifiability of β with respect to the norm $\|\cdot\|_I$.

Comment 21 (A comment on the identifiability factor $\kappa_I(\beta)$). *The identifiability factors $\kappa_I(\beta)$ depend on the true parameter value β. This is not the main focus of this section, but we note that these factors represent a modest generalization of the cone invertibility factors and sensitivity characteristics defined in [114] and [48], which are known to be quite general. The main difference perhaps is the use of a norm of interest $\|\cdot\|_I$ instead of the ℓ_q norms and the use of smaller (non-conic) restricted set $\mathcal{R}(\beta)$ in the definition. It is useful to note for later comparisons that in the case of prediction norm $\|\cdot\|_I = \|\cdot\|_{\mathrm{pr}}$ and under the exact sparsity assumption $\|\beta\|_0 \leqslant s$, we have*

$$\kappa_{\mathrm{pr}}(\beta) \geqslant 2^{-1}s^{-1/2}\kappa(s,1), \tag{3.17}$$

where $\kappa(s,1)$ is the restricted eigenvalue defined in [19]. □

Next we state bounds on the estimation error for the Dantzig selector $\widehat{\beta}^{(0)}$ with canonical penalty level $\lambda = \lambda^{(0)} := c_0(1-\alpha)$ and the Dantzig selector $\widehat{\beta}^{(1)}$ with design-adaptive penalty level $\lambda = \lambda^{(1)} := c_{Z_0}(1-\alpha)$.

Theorem 17 (Performance of Dantzig Selector in Non-Gaussian Model). *Suppose that there are some constants $c_1 > 0, C_1 > 0$ and $\sigma^2 > 0$, and a sequence $B_n \geqslant 1$ of constants such that for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$: (i) $|z_{ij}| \leqslant B_n$; (ii) $\mathbb{E}_n[z_{ij}^2] = 1$; (iii) $\mathrm{E}[\varepsilon_i^2] = \sigma^2$; (iv) $\mathrm{E}[\varepsilon_i^4] \leqslant C_1$; and (v) $B_n^4(\log(pn))^7/n \leqslant C_1 n^{-c_1}$. Then there exist constants $c > 0$ and $C > 0$ depending only on c_1, C_1 and σ^2 such that, with probability at least $1-\alpha-Cn^{-c}$, for either $k = 0$ or 1,*

$$\|\widehat{\beta}^{(k)} - \beta\|_I \leqslant \frac{2\lambda^{(k)}}{\sqrt{n}\kappa_I(\beta)}.$$

The most important feature of this result is that it provides Gaussian-like conclusions (as explained below) in a model with non-Gaussian noise, having only four bounded moments. However, the probabilistic guarantee is not $1-\alpha$ as, e.g., in [19], but rather $1-\alpha-Cn^{-c}$, which reflects the cost of non-Gaussianity (along with more stringent side conditions). In what follows we discuss details of this result. Note that the bound above holds for any semi-norm of interest $\|\cdot\|_I$.

Comment 22 (Improved Performance from Design-Adaptive Penalty Level). *The use of the design-adaptive penalty level implies a better performance guarantee for $\widehat{\beta}^{(1)}$ over $\widehat{\beta}^{(0)}$. Indeed, we have*

$$\frac{2c_{Z_0}(1-\alpha)}{\sqrt{n}\kappa_I(\beta)} \leqslant \frac{2c_0(1-\alpha)}{\sqrt{n}\kappa_I(\beta)}.$$

E.g., in some designs, we can have $\sqrt{n}\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}e_i]| = O_{\mathrm{P}}(1)$, so that $c_{Z_0}(1-\alpha) = O(1)$, whereas $c_0(1-\alpha) \propto \sqrt{\log p}$. Thus, the performance guarantee provided by $\widehat{\beta}^{(1)}$ can be much better than that of $\widehat{\beta}^{(0)}$. □

Comment 23 (Relation to the previous results under Gaussianity). *To compare to the previous results obtained for the Gaussian settings, let us focus on the prediction norm and on estimator $\widehat{\beta}^{(1)}$ with penalty level $\lambda = c_{Z_0}(1-\alpha)$. Suppose that the true* *value β is sparse, namely $\|\beta\|_0 \leqslant s$. In this case, with probability at least $1-\alpha-Cn^{-c}$,*

$$\|\widehat{\beta}^{(1)} - \beta\|_{\rm pr} \leqslant \frac{2c_{Z_0}(1-\alpha)}{\sqrt{n}\kappa_{\rm pr}(\beta)} \leqslant \frac{4\sqrt{s}c_0(1-\alpha)}{\sqrt{n}\kappa(s,1)} \leqslant \frac{4\sqrt{s}\sqrt{2\log(\alpha/(2p))}}{\sqrt{n}\kappa(s,1)}, \tag{3.18}$$

where the last bound is the same as in [19], Theorem 7.1, obtained for the Gaussian case. We recover the same (or tighter) upper bound without making the Gaussianity assumption on the errors. However, the probabilistic guarantee is not $1-\alpha$ as in [19], but rather $1-\alpha-Cn^{-c}$, which together with side conditions is the cost of non-Gaussianity. □

Comment 24 (Other refinements). *Unrelated to the main theme of this paper, we can see from (3.18) that there is some tightening of the performance bound due to the use of the identifiability factor $\kappa_{\rm pr}(\beta)$ in place of the restricted eigenvalue $\kappa(s,1)$; for example, if $p=2$ and $s=1$ and the two regressors are identical, then $\kappa_{\rm pr}(\beta)>0$, whereas $\kappa(1,1)=0$. There is also some tightening due to the use of $c_{Z_0}(1-\alpha)$ instead of $c_0(1-\alpha)$ as penalty level, as mentioned above.* □

3.4.2 Heteroscedastic case.

We consider the same model as above, except now the assumption on the error becomes

$$\sigma_i^2 := \mathrm{E}[\varepsilon_i^2] \leqslant \sigma^2, \quad i=1,\ldots,n,$$

i.e., σ^2 is the upper bound on the conditional variance, and we assume that this bound is known (for simplicity). As before, ideally we would like to set penalty level λ equal to

$$c_{T_0}(1-\alpha) := (1-\alpha)\text{-quantile of } T_0,$$

(where T_0 is defined above, and we note that z_i are treated as fixed). The CLT applies as before, namely the difference of the distribution functions of T_0 and its Gaussian analogue Z_0 converges to zero. In this case, the Gaussian analogue can be represented as

$$Z_0 := \sqrt{n}\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}\sigma_i e_i]|.$$

Unlike in the homoscedastic case, the covariance structure is no longer known, since σ_i are unknown and we can no longer calculate the quantiles of Z_0. However, we can estimate them using the following multiplier bootstrap procedure.

First, we estimate the residuals $\widehat{\varepsilon}_i = y_i - z_i'\widehat{\beta}^{(0)}$ obtained from a preliminary Dantzig selector $\widehat{\beta}^{(0)}$ with the conservative penalty level $\lambda = \lambda^{(0)} := c_0(1-1/n) := \sigma\Phi^{-1}(1-1/(2pn))$, where σ^2 is the upper bound on the error variance assumed to be known. Let $(e_i)_{i=1}^n$ be a sequence of i.i.d. standard Gaussian random variables, and let

$$W := \sqrt{n}\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}\widehat{\varepsilon}_i e_i]|.$$

Then we estimate $c_{Z_0}(1-\alpha)$ by

$$c_W(1-\alpha) := (1-\alpha)\text{-quantile of } W,$$

defined conditional on data $(z_i, y_i)_{i=1}^n$. Note that $c_W(1-\alpha)$ can be calculated numerically with any specified precision by the simulation. Then we apply program (3.16) with $\lambda = \lambda^{(1)} = c_W(1-\alpha)$ to obtain $\widehat{\beta}^{(1)}$.

Theorem 18 (Performance of Dantzig in Non-Gaussian Model with Bootstrap Penalty Level). *Suppose that there are some constants $c_1 > 0, C_1 > 0, \underline{\sigma}^2 > 0$ and $\sigma^2 > 0$, and a sequence $B_n \geqslant 1$ of constants such that for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$: (i) $|z_{ij}| \leqslant B_n$; (ii) $\mathbb{E}_n[z_{ij}^2] = 1$; (iii) $\underline{\sigma}^2 \leqslant \mathrm{E}[\varepsilon_i^2] \leqslant \sigma^2$; (iv) $\mathrm{E}[\varepsilon_i^4] \leqslant C_1$; (v) $B_n^4(\log(pn))^7/n \leqslant C_1 n^{-c_1}$; and (vi) $(\log p)B_n c_0(1-1/n)/(\sqrt{n}\kappa_{\mathrm{pr}}(\beta)) \leqslant C_1 n^{-c_1}$. Then there exist constants $c > 0$ and $C > 0$ depending only on $c_1, C_1, \underline{\sigma}^2$ and σ^2 such that, with probability at least $1-\alpha-\nu_n$ where $\nu_n = Cn^{-c}$, we have*

$$\|\widehat{\beta}^{(1)} - \beta\|_I \leqslant \frac{2\lambda^{(1)}}{\sqrt{n}\kappa_I(\beta)}. \tag{3.19}$$

Moreover, with probability at least $1-\nu_n$,

$$\lambda^{(1)} = c_W(1-\alpha) \leqslant c_{Z_0}(1-\alpha+\nu_n),$$

where $c_{Z_0}(1-a) :=$ $(1-a)$*-quantile of* Z_0*; in particular* $c_{Z_0}(1-a) \leqslant c_0(1-a)$.

3.4.3 Some Extensions

Here we comment on some additional potential applications.

Comment 25 (Confidence Sets). *Note that bounds given in the preceding theorems can be used for inference on* β *or components of* β*, given the assumption* $\kappa_I(\beta) \geqslant \kappa$*, where* κ *is a known constant. For example, consider inference on the* j*-th component* β_j *of* β*. In this case, we take the norm of interest* $\|\delta\|_I$ *to be* $\|\delta\|_{jc} = |\delta_j|$ *on* $\mathbb{R}^p$*, and consider the corresponding identifiability factor* $\kappa_{jc}(\beta)$*. Suppose it is known that* $\kappa_{jc}(\beta) \geqslant \kappa$*. Then a* $(1-\alpha-Cn^{-c})$*-confidence interval for* β_j *is given by*

$$\{b \in \mathbb{R} : |\widehat{\beta}_j^{(1)} - b| \leqslant 2\lambda^{(1)}/(\sqrt{n}\kappa)\}.$$

This confidence set is of interest, but it does require the investigator to make a stance on what a plausible κ *should be. We refer to [48] for a justification of confidence sets of this type and possible ways of computing lower bounds on* κ*; there is also a work by [64], which provides computable lower bounds on related quantities.* □

Comment 26 (Generalization of Dantzig Selector). *There are many interesting applications where the results given above apply. There are, for example, interesting works by [1] and [47] that consider related estimators that minimize a convex penalty subject to the multiresolution screening constraints. In the context of the regression problem studied above, such estimators may be defined as:*

$$\widehat{\beta} \in \arg\min_{b \in \mathbb{R}^p} J(b) \text{ subject to } \sqrt{n} \max_{1 \leqslant j \leqslant p} |\mathbb{E}_n[z_{ij}(y_i - z_i'b)]| \leqslant \lambda,$$

where J *is a convex penalty, and the constraint is used for multiresolution screening. For example, the Lasso estimator is nested by the above formulation by using* $J(b) = \|b\|_{\mathrm{pr}}$*, and the previous Dantzig selector by using* $J(b) = \|b\|_{\ell_1}$*; the estimators can be interpreted as a point in confidence set for* β*, which lies closest to zero under* J*-discrepancy (see references above for both of these points). Our results on choosing* λ

apply to this class of estimators, and the previous analysis also applies by redefining the identifiability factor $\kappa_I(\beta)$ relative to the new restricted set $\mathcal{R}(\beta) := \{\delta \in \mathbb{R}^p : J(\beta+\delta) \leqslant J(\beta)\}$; where $\kappa_I(\beta)$ is defined as ∞ if $\mathcal{R}(\beta) = \{0\}$. □

3.5 Appendix A. Preliminaries

3.5.1 A Useful Maximal Inequality

The following lemma, which is derived in [34], is a useful variation of standard maximal inequalities.

Lemma 37 (Maximal Inequality). *Let $x_1, \dots, x_n$ be independent random vectors in $\mathbb{R}^p$ with $p \geqslant 2$. Let $M = \max_{1\leqslant i \leqslant n} \max_{1 \leqslant j \leqslant p} |x_{ij}|$ and $\sigma^2 = \max_{1\leqslant j\leqslant p} \bar{\mathrm{E}}[x_{ij}^2]$. Then*

$$\mathrm{E}\left[\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[x_{ij}] - \bar{\mathrm{E}}[x_{ij}]|\right] \lesssim \sigma\sqrt{(\log p)/n} + \sqrt{\mathrm{E}[M^2]}(\log p)/n.$$

Proof. See [34], Lemma 8. □

3.5.2 Properties of the Smooth Max Function

We will use the following properties of the smooth max function.

Lemma 38 (Properties of F_β). *For every $1 \leqslant j,k,l \leqslant p$,*

$$\partial_j F_\beta(z) = \pi_j(z), \quad \partial_j\partial_k F_\beta(z) = \beta w_{jk}(z), \quad \partial_j\partial_k\partial_l F_\beta(z) = \beta^2 q_{jkl}(z).$$

where, for $\delta_{jk} := 1\{j = k\}$,

$$\pi_j(z) := e^{\beta z_j}/\textstyle\sum_{m=1}^p e^{\beta z_m}, \; w_{jk}(z) := (\pi_j\delta_{jk} - \pi_j\pi_k)(z),$$
$$q_{jkl}(z) := (\pi_j\delta_{jl}\delta_{jk} - \pi_j\pi_l\delta_{jk} - \pi_j\pi_k(\delta_{jl}+\delta_{kl}) + 2\pi_j\pi_k\pi_l)(z).$$

Moreover,

$$\pi_j(z) \geqslant 0, \; \textstyle\sum_{j=1}^p \pi_j(z) = 1, \; \sum_{j,k=1}^p |w_{jk}(z)| \leqslant 2, \; \sum_{j,k,l=1}^p |q_{jkl}(z)| \leqslant 6.$$

Proof of Lemma 38. The first property was noted in [28]. The other properties follow from repeated application of the chain rule. □

Lemma 39 (Lipschitz Property of F_β). *For every $x \in \mathbb{R}^p$ and $z \in \mathbb{R}^p$, we have $|F_\beta(x) - F_\beta(z)| \leqslant \max_{1 \leqslant j \leqslant p} |x_j - z_j|$.*

Proof of Lemma 39. For some $t \in [0,1]$,

$$\begin{aligned} |F_\beta(x) - F_\beta(z)| &= |\textstyle\sum_{j=1}^p \partial_j F_\beta(x + t(z-x))(z_j - x_j)| \\ &\leqslant \textstyle\sum_{j=1}^p \pi_j(x + t(z-x)) \displaystyle\max_{1 \leqslant j \leqslant p} |z_j - x_j| \leqslant \max_{1 \leqslant j \leqslant p} |z_j - x_j|, \end{aligned}$$

where the property $\sum_{j=1}^p \pi_j(x + t(z-x)) = 1$ was used. □

We will also use the following properties of $m = g \circ F_\beta$. Here we assume $g \in C_b^3(\mathbb{R})$ in Lemmas 40-42 below.

Lemma 40 (Three derivatives of $m = g \circ F_\beta$). *For every $1 \leqslant j,k,l \leqslant p$,*

$$\begin{aligned} \partial_j m(z) &= (\partial g(F_\beta)\pi_j)(z), \\ \partial_j \partial_k m(z) &= (\partial^2 g(F_\beta)\pi_j\pi_k + \partial g(F_\beta)\beta w_{jk})(z), \\ \partial_j \partial_k \partial_l m(z) &= (\partial^3 g(F_\beta)\pi_j\pi_k\pi_l + \partial^2 g(F_\beta)\beta(w_{jk}\pi_l + w_{jl}\pi_k + w_{kl}\pi_j) \\ &\quad + \partial g(F_\beta)\beta^2 q_{jkl})(z), \end{aligned}$$

where π_j, w_{jk} and q_{jkl} are defined in Lemma 38, and (z) denotes evaluation at z, including evaluation of F_β at z.

Proof of lemma 40. The proof follows from repeated application of the chain rule and by the properties noted in Lemma 38. □

Lemma 41 (Bounds on derivatives of $m = g \circ F_\beta$). *For every $1 \leqslant j,k,l \leqslant p$,*

$$|\partial_j \partial_k m(z)| \leqslant U_{jk}(z), \quad |\partial_j \partial_k \partial_l m(z)| \leqslant U_{jkl}(z),$$

where

$$U_{jk}(z) := (G_2\pi_j\pi_k + G_1\beta W_{jk})(z),\ W_{jk}(z) := (\pi_j\delta_{jk} + \pi_j\pi_k)(z),$$
$$U_{jkl}(z) := (G_3\pi_j\pi_k\pi_l + G_2\beta(W_{jk}\pi_l + W_{jl}\pi_k + W_{kl}\pi_j) + G_1\beta^2Q_{jkl})(z),$$
$$Q_{jkl}(z) := (\pi_j\delta_{jl}\delta_{jk} + \pi_j\pi_l\delta_{jk} + \pi_j\pi_k(\delta_{jl} + \delta_{kl}) + 2\pi_j\pi_k\pi_l)(z).$$

Moreover,

$$\textstyle\sum_{j,k=1}^{p} U_{jk}(z) \leqslant (G_2 + 2G_1\beta),\ \sum_{j,k,l=1}^{p} U_{jkl}(z) \leqslant (G_3 + 6G_2\beta + 6G_1\beta^2).$$

Proof of Lemma 41. The lemma follows from a direct calculation. □

Lemma 42 (Stability). *For every $z \in \mathbb{R}^p$, $w \in \mathbb{R}^p$ such that* $\max_{j\leqslant p}|w_j|\beta \leqslant 1$, $\tau \in [0,1]$, *and every* $1 \leqslant j,k,l \leqslant p$, *we have*

$$U_{jk}(z) \lesssim U_{jk}(z+\tau w) \lesssim U_{jk}(z),\ U_{jkl}(z) \lesssim U_{jkl}(z+\tau w) \lesssim U_{jkl}(z).$$

Proof of Lemma 42. Observe that

$$\pi_j(z+\tau w) = \frac{e^{z_j\beta+\tau w_j\beta}}{\sum_{m=1}^{p} e^{z_m\beta+\tau w_m\beta}} \leqslant \frac{e^{z_j\beta}}{\sum_{m=1}^{p} e^{z_m\beta}} \cdot \frac{e^{\tau \max_{j\leqslant p}|w_j|\beta}}{e^{-\tau \max_{j\leqslant p}|w_j|\beta}} \leqslant e^2\pi_j(z).$$

Similarly, $\pi_j(z+\tau w) \geqslant e^{-2}\pi_j(z)$. Since U_{jk} and U_{jkl} are finite sums of products of terms such as π_j, π_k, π_l, δ_{jk}, the claim of the lemma follows. □

3.5.3 Lemma on Truncation

The proof of Theorem 14 uses the following properties of the truncation operation. Recall that $\tilde{x}_i = (\tilde{x}_{ij})_{j=1}^p$ and $\tilde{X} = n^{-1/2}\sum_{i=1}^n \tilde{x}_i$, where "tilde" denotes the truncation operation defined in Section 2. The following lemma also covers the special case where $(x_i)_{i=1}^n = (y_i)_{i=1}^n$. The property (d) is a consequence of sub-Gaussian inequality of [39], Theorem 2.16. for self-normalized sums.

Lemma 43 (Truncation Impact). *For every* $1 \leqslant j,k \leqslant p$ *and* $q \geqslant 1$, *(a)* $(\bar{\mathrm{E}}[|\tilde{x}_{ij}|^q])^{1/q} \leqslant$

$2(\bar{\mathrm{E}}[|x_{ij}|^q])^{1/q}$; *(b)* $\bar{\mathrm{E}}[|\tilde{x}_{ij}\tilde{x}_{ik} - x_{ij}x_{ik}|] \leqslant (3/2)(\bar{\mathrm{E}}[x_{ij}^2] + \bar{\mathrm{E}}[x_{ik}^2])\varphi(u)$; *(c)* $\mathbb{E}_n[(\mathrm{E}[x_{ij} - \tilde{x}_{ij}])^2] \leqslant \bar{\mathrm{E}}[(x_{ij} - \tilde{x}_{ij})^2] \leqslant \bar{\mathrm{E}}[x_{ij}^2]\varphi^2(u)$. *Moreover, for a given* $\gamma \in (0,1)$, *let* $u \geqslant u(\gamma)$ *where* $u(\gamma)$ *is defined in Section 3.2. Then: (d) with probability at least* $1 - 5\gamma$, *for all* $1 \leqslant j \leqslant p$,

$$|X_j - \tilde{X}_j| \leqslant 5\sqrt{\bar{\mathrm{E}}[x_{ij}^2]}\varphi(u)\sqrt{2\log(p/\gamma)}.$$

Proof. See Appendix 3.10. □

3.6 Appendix B. Proofs for Section 2

3.6.1 Proof of Theorem 13

Recall that we are assuming that sequences $(x_i)_{i=1}^n$ and $(y_i)_{i=1}^n$ are independent. For $t \in [0,1]$, we consider the Slepian interpolation between Y and X:

$$Z(t) := \sqrt{t}X + \sqrt{1-t}Y = \sum_{i=1}^n Z_i(t),\ Z_i(t) := \frac{1}{\sqrt{n}}(\sqrt{t}x_i + \sqrt{1-t}y_i).$$

We shall also employ Stein's leave-one-out expansions:

$$Z^{(i)}(t) := (Z_{ij}(t))_{j=1}^p := Z(t) - Z_i(t).$$

Let $\Psi(t) = \mathrm{E}[m(Z(t))]$ for $m := g \circ F_\beta$. Then by Taylor's theorem,

$$\begin{aligned}\mathrm{E}[m(X) - m(Y)] &= \Psi(1) - \Psi(0) = \int_0^1 \Psi'(t)dt \\ &= \frac{1}{2}\sum_{j=1}^p\sum_{i=1}^n \int_0^1 \mathrm{E}[\partial_j m(Z(t))\dot{Z}_{ij}(t)]dt = \frac{1}{2}(I + II + III),\end{aligned}$$

where

$$\begin{aligned}
\dot{Z}_{ij}(t) &= \frac{d}{dt}Z_{ij}(t) = \frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{t}}x_{ij} - \frac{1}{\sqrt{1-t}}y_{ij}\right), \text{ and} \\
I &= \sum_{j=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[\partial_j m(Z^{(i)}(t))\dot{Z}_{ij}(t)]dt, \\
II &= \sum_{j,k=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[\partial_j\partial_k m(Z^{(i)}(t))\dot{Z}_{ij}(t)Z_{ik}(t)]dt, \\
III &= \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1\int_0^1 (1-\tau)\mathrm{E}[\partial_j\partial_k\partial_l m(Z^{(i)}(t)+\tau Z_i(t))\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)]d\tau dt.
\end{aligned}$$

Note that random variable $Z^{(i)}(t)$ and random vector $(\dot{Z}_{ij}(t), Z_{ij}(t))$ are independent, and $\mathrm{E}[\dot{Z}_{ij}(t)] = 0$. Hence we have $I = 0$; moreover, since $\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)] = n^{-1}\mathrm{E}[x_{ij}x_{ik} - y_{ij}y_{ik}] = 0$ by construction of $(y_i)_{i=1}^{n}$, we also have $II = 0$. Consider the third term III. We have that

$$\begin{aligned}
|III| &\lesssim_{(1)} (G_3 + G_2\beta + G_1\beta^2)n\int \mathrm{E}\left[\max_{1\leqslant j,k,l\leqslant p}|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|\right]dt, \\
&\lesssim_{(2)} n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)\bar{\mathrm{E}}\left[\max_{1\leqslant j\leqslant p}(|x_{ij}| + |y_{ij}|)^3\right],
\end{aligned}$$

where (1) follows from $|\partial_j\partial_k\partial_l m(Z^{(i)}(t) + \tau Z_i(t))| \leqslant U_{jkl}(Z^{(i)}(t) + \tau Z_i(t)) \lesssim (G_3 + G_2\beta + G_1\beta^2)$ holding by Lemma 41, and (2) is shown below. The first claim of the theorem now follows. The second claim follows directly from property (3.9) of the smooth max function.

It remains to show (2). Define $\omega(t) = 1/(\sqrt{t}\wedge\sqrt{1-t})$ and note,

$$\begin{aligned}
&\int_0^1 n\bar{\mathrm{E}}\left[\max_{1\leqslant j,k,l\leqslant p}|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|\right]dt \\
&= \int_0^1 \omega(t)n\bar{\mathrm{E}}\left[\max_{1\leqslant j,k,l\leqslant p}|\dot{Z}_{ij}(t)/\omega(t))Z_{ik}(t)Z_{il}(t)|\right]dt \\
&\leqslant n\int_0^1 \omega(t)\left(\bar{\mathrm{E}}[\max_{1\leqslant j\leqslant p}|\dot{Z}_{ij}(t)/\omega(t)|^3]\bar{\mathrm{E}}[\max_{1\leqslant j\leqslant p}|Z_{ij}(t)|^3]\bar{\mathrm{E}}[\max_{1\leqslant j\leqslant p}|Z_{ij}(t)|^3]\right)^{1/3}dt \\
&\leqslant n^{-1/2}\left\{\int_0^1 \omega(t)dt\right\}\bar{\mathrm{E}}\left[\max_{1\leqslant j\leqslant p}(|x_{ij}| + |y_{ij}|)^3\right]
\end{aligned}$$

where the first inequality follows from Hölder's inequality, and the second from the fact that $|\dot{Z}_{ij}(t)/\omega(t)| \leqslant (|x_{ij}|+|y_{ij}|)/\sqrt{n}$, $|Z_{ij}(t)| \leqslant (|x_{ik}|+|y_{ik}|)/\sqrt{n}$. Finally we note that $\int_0^1 \omega(t)dt \lesssim 1$, so inequality (2) follows. This completes the overall proof. $\square$

3.6.2 Proof of Corollary 3

In this proof, let $C > 0$ denote a generic constant depending only on c_1 and C_1, and its value may change from place to place. For $\beta > 0$, define $e_\beta := \beta^{-1}\log p$. Recall that $S_i := \max_{1 \leqslant j \leqslant p}(|x_{ij}| + |y_{ij}|)$. Consider and fix a C^3-function $g_0 : \mathbb{R} \to [0,1]$ such that $g_0(s) = 1$ for $s \leqslant 0$ and $g_0(s) = 0$ for $s \geqslant 1$. Fix any $t \in \mathbb{R}$, and define $g(s) = g_0(\psi(s - t - e_\beta))$. For this function g, $G_0 = 1$, $G_1 \lesssim \psi$, $G_2 \lesssim \psi^2$ and $G_3 \lesssim \psi^3$.

Observe now that

$$\begin{aligned}
\mathrm{P}(T_0 \leqslant t) &\leqslant \mathrm{P}(F_\beta(X) \leqslant t + e_\beta) \leqslant \mathrm{E}[g(F_\beta(X))] \\
&\leqslant \mathrm{E}[g(F_\beta(Y))] + C(\psi^3 + \beta\psi^2 + \beta^2\psi)(n^{-1/2}\bar{\mathrm{E}}[S_i^3]) \\
&\leqslant \mathrm{P}(F_\beta(Y) \leqslant t + e_\beta + \psi^{-1}) + C(\psi^3 + \beta\psi^2 + \beta^2\psi)(n^{-1/2}\bar{\mathrm{E}}[S_i^3]) \\
&\leqslant \mathrm{P}(Z_0 \leqslant t + e_\beta + \psi^{-1}) + C(\psi^3 + \beta\psi^2 + \beta^2\psi)(n^{-1/2}\bar{\mathrm{E}}[S_i^3]).
\end{aligned}$$

where the first inequality follows from (3.9), the second from construction of g, the third from Theorem 13, and the fourth from construction of g, and the last from (3.9). The remaining step is to compare $\mathrm{P}(Z_0 \leqslant t + e_\beta + \psi^{-1})$ with $\mathrm{P}(Z_0 \leqslant t)$ and this is where Lemma 32 plays its role. By Lemma 32,

$$\mathrm{P}(Z_0 \leqslant t + e_\beta + \psi^{-1}) - \mathrm{P}(Z_0 \leqslant t) \leqslant C(e_\beta + \psi^{-1})\sqrt{1 \vee \log(p\psi)}.$$

by which we have

$$\mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t) \leqslant C[(\psi^3 + \beta\psi^2 + \beta^2\psi)(n^{-1/2}\bar{\mathrm{E}}[S_i^3]) + (e_\beta + \psi^{-1})\sqrt{1 \vee \log(p\psi)}].$$

We have to minimize the right side with respect to β and ψ. It is reasonable to choose β in such a way that e_β and ψ^{-1} are balanced, i.e., $\beta = \psi \log p$. With this β, the bracket on the right side is

$$\lesssim \psi^3(\log p)^2(n^{-1/2}\bar{\mathrm{E}}[S_i^3]) + \psi^{-1}\sqrt{1 \vee \log(p\psi)},$$

which is approximately minimized by $\psi = (\log p)^{-3/8}(n^{-1/2}\bar{\mathrm{E}}[S_i^3])^{-1/4}$. With this ψ, $\psi \leqslant (n^{-1/2}\bar{\mathrm{E}}[S_i^3])^{-1/4} \leqslant Cn^{1/8}$ (recall that $p \geqslant 3$), and hence $\log(p\psi) \leqslant C\log(pn)$. Therefore,

$$\mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t) \leqslant C(n^{-1/2}\bar{\mathrm{E}}[S_i^3])^{1/4}(\log(pn))^{7/8}.$$

This gives one half of the claim. The other half follows similarly. □

3.6.3 Proof of Theorem 14

The second claim of the theorem follows from property (3.9) of the smooth max function. Hence we shall prove the first claim. The proof strategy is similar to the proof of Theorem 13. However, to control effectively the third order terms in the leave-one-out expansions we shall use truncation and replace X and Y by their truncated versions $\tilde{X}$ and $\tilde{Y}$, defined as follows: let $\tilde{x}_i = (\tilde{x}_{ij})_{j=1}^p$, where $\tilde{x}_{ij}$ was defined before the statement of the theorem, and define the truncated version of X as $\tilde{X} = n^{-1/2}\sum_{i=1}^n \tilde{x}_i$. Also let

$$\tilde{y}_i := (\tilde{y}_{ij})_{j=1}^p,\ \tilde{y}_{ij} := y_{ij}1\left\{|y_{ij}| \leqslant u(\bar{\mathrm{E}}[y_{ij}^2])^{1/2}\right\},\ \tilde{Y} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{y}_i.$$

Note that by the symmetry of the distribution of y_{ij}, $\mathrm{E}[\tilde{y}_{ij}] = 0$. Recall that we are assuming that sequences $(x_i)_{i=1}^n$ and $(y_i)_{i=1}^n$ are independent.

The proof consists of four steps. Step 1 will show that we can replace X by $\tilde{X}$ and Y by $\tilde{Y}$. Step 2 will bound the difference of the expectations of the relevant functions of $\tilde{X}$ and $\tilde{Y}$. This is the main step of the proof. Steps 3 and 4 will carry out supporting calculations. The steps of the proof will also call on various technical lemmas collected in Appendix 3.5.

Step 1. Let $m := g \circ F_\beta$. The main goal is to bound $\mathrm{E}[m(X) - m(Y)]$. Define

$$\mathcal{I} = 1\left\{ \max_{1 \leqslant j \leqslant p} |X_j - \tilde{X}_j| \leqslant \Delta(\gamma, u) \text{ and } \max_{1 \leqslant j \leqslant p} |Y_j - \tilde{Y}_j| \leqslant \Delta(\gamma, u) \right\},$$

where $\Delta(\gamma, u) := 5M_2\varphi(u)\sqrt{2\log(p/\gamma)}$. By Lemma 43 we have $\mathrm{E}[\mathcal{I}] \geqslant 1 - 10\gamma$. Observe that by Lemma 39,

$$|m(x) - m(y)| \leqslant G_1|F_\beta(x) - F_\beta(y)| \leqslant G_1 \max_{1 \leqslant j \leqslant p} |x_j - y_j|,$$

so that

$$\begin{aligned} |\mathrm{E}[m(X) - m(\tilde{X})]| &\leqslant |\mathrm{E}[(m(X) - m(\tilde{X}))\mathcal{I}]| + |\mathrm{E}[(m(X) - m(\tilde{X}))(1 - \mathcal{I})]| \\ &\lesssim G_1\Delta(\gamma, u) + G_0\gamma, \\ |\mathrm{E}[m(Y) - m(\tilde{Y})]| &\leqslant \mathrm{E}[(m(Y) - m(\tilde{Y}))\mathcal{I}]| + |\mathrm{E}[(m(Y) - m(\tilde{Y}))(1 - \mathcal{I})] \\ &\lesssim G_1\Delta(\gamma, u) + G_0\gamma, \end{aligned}$$

and hence

$$|\mathrm{E}[m(X) - m(Y)]| \lesssim |\mathrm{E}[m(\tilde{X}) - m(\tilde{Y})]| + G_1\Delta(\gamma, u) + G_0\gamma.$$

Step 2. (Main Step) The purpose of this step is to establish the bound:

$$|\mathrm{E}[m(\tilde{X}) - m(\tilde{Y})]| \lesssim n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)M_3^3 + (G_2 + \beta G_1)M_2^2\varphi(u).$$

Define, as in the proof of Theorem 13,

$$\begin{aligned} &Z(t) := \sqrt{t}\tilde{X} + \sqrt{1-t}\tilde{Y} = \sum_{i=1}^{n} Z_i(t), \quad Z_i(t) := \frac{1}{\sqrt{n}}(\sqrt{t}\tilde{x}_i + \sqrt{1-t}\tilde{y}_i), \text{ and} \\ &Z^{(i)}(t) := Z(t) - Z_i(t), \ \dot{Z}_{ij}(t) = \frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{t}}\tilde{x}_{ij} - \frac{1}{\sqrt{1-t}}\tilde{y}_{ij}\right). \end{aligned}$$

Arguing as in the proof of Theorem 13, we have

$$\mathrm{E}[m(\tilde{X}) - m(\tilde{Y})] = \frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[\partial_j m(Z(t))\dot{Z}_{ij}(t)]dt = \frac{1}{2}(I + II + III),$$

where

$$\begin{aligned} I &= \sum_{j=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[\partial_j m(Z^{(i)}(t))\dot{Z}_{ij}(t)]dt, \\ II &= \sum_{j,k=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[\partial_j\partial_k m(Z^{(i)}(t))\dot{Z}_{ij}(t)Z_{ik}(t)]dt, \\ III &= \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1\int_0^1 (1-\tau)\mathrm{E}[\partial_j\partial_k\partial_l m(Z^{(i)}(t) + \tau Z_i(t))\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)]d\tau dt. \end{aligned}$$

By independence of $Z^{(i)}(t)$ and $\dot{Z}_{ij}(t)$ together with the fact that $\mathrm{E}[\dot{Z}_{ij}(t)] = 0$, we have $I = 0$. Moreover, in steps 3 and 4 below, we will show that

$$|II| \lesssim (G_2 + \beta G_1)M_2^2\varphi(u), \ |III| \lesssim n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)M_3^3.$$

The claim of this step now follows.

Step 3. (Bound on II) By independence of $Z^{(i)}(t)$ and $\dot{Z}_{ij}(t)Z_{ik}(t)$,

$$\begin{aligned} |II| &= \left|\sum_{j,k=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[\partial_j\partial_k m(Z^{(i)}(t))]\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]dt\right| \\ &\leqslant \sum_{j,k=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[|\partial_j\partial_k m(Z^{(i)}(t))|]\cdot|\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]|dt \\ &\leqslant \sum_{j,k=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[U_{jk}(Z^{(i)}(t))] \ \cdot|\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]|dt, \end{aligned}$$

where the last step follows from Lemma 41. Since $|\sqrt{t}\tilde{x}_{ij} + \sqrt{1-t}\tilde{y}_{ij}| \leqslant 2\sqrt{2}uM_2$, so that $|\beta(\sqrt{t}\tilde{x}_{ij}+\sqrt{1-t}\tilde{y}_{ij})/\sqrt{n}| \leqslant 1$ (which is satisfied by the assumption $\beta 2\sqrt{2}uM_2/\sqrt{n} \leqslant$

1), by Lemmas 42 and 41, the last expression is bounded by

$$\begin{aligned}
&\sum_{j,k=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[U_{jk}(Z(t))]\cdot|\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]|dt\\
&=\int_0^1\left\{\sum_{j,k=1}^{p}\mathrm{E}[U_{jk}(Z(t))]\right\}\sum_{i=1}^{n}|\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]|dt\\
&\lesssim (G_2+G_1\beta)\int_0^1\sum_{i=1}^{n}|\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]|dt.
\end{aligned}$$

Observe that since $\mathrm{E}[x_{ij}x_{ik}]=\mathrm{E}[y_{ij}y_{ik}]$, we have that $\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]=n^{-1}\mathrm{E}[\tilde{x}_{ij}\tilde{x}_{ik}-\tilde{y}_{ij}\tilde{y}_{ik}]=n^{-1}\mathrm{E}[\tilde{x}_{ij}\tilde{x}_{ik}-x_{ij}x_{ik}]+n^{-1}\mathrm{E}[y_{ij}y_{ik}-\tilde{y}_{ij}\tilde{y}_{ik}]$, so that by Lemma 43 (b), $\sum_{i=1}^{n}|\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]|\leqslant\bar{\mathrm{E}}[|\tilde{x}_{ij}\tilde{x}_{ik}-x_{ij}x_{ik}|]+\bar{\mathrm{E}}[|y_{ij}y_{ik}-\tilde{y}_{ij}\tilde{y}_{ik}|]\lesssim(\bar{\mathrm{E}}[x_{ij}^2]+\bar{\mathrm{E}}[x_{ik}^2])\varphi(u)\lesssim M_2^2\varphi(u)$. Therefore, we conclude that $|II|\lesssim(G_2+G_1\beta)M_2^2\varphi(u)$.

Step 4. (Bound on III) Observe that

$$\begin{aligned}
|III| &\leqslant_{(1)} \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1\int_0^1 \mathrm{E}[U_{jkl}(Z^{(i)}(t)+\tau Z_i(t))|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]d\tau dt\\
&\lesssim_{(2)} \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[U_{jkl}(Z^{(i)}(t))|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt\\
&=_{(3)} \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[U_{jkl}(Z^{(i)}(t))]\cdot\mathrm{E}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt, \qquad (3.20)
\end{aligned}$$

where (1) follows from $|\partial_j\partial_k\partial_l m(z)|\leqslant U_{jkl}(z)$ (see Lemma 41), (2) from Lemma 42, (3) from independence of $Z^{(i)}(t)$ and $\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)$. Moreover, the last expression is bounded as follows:

$$
\begin{aligned}
&\text{right side of (3.20)} \lesssim_{(4)} \sum_{j,k,l=1}^{p} \sum_{i=1}^{n} \int_0^1 \mathrm{E}[U_{jkl}(Z(t))] \cdot \mathrm{E}[|\dot{Z}_{ij}(t) Z_{ik}(t) Z_{il}(t)|]dt \\
&=_{(5)} \sum_{j,k,l=1}^{p} \int_0^1 \mathrm{E}[U_{jkl}(Z(t))] \cdot n\bar{\mathrm{E}}[|\dot{Z}_{ij}(t) Z_{ik}(t) Z_{il}(t)|]dt \\
&\leqslant_{(6)} \int_0^1 \left(\sum_{j,k,l=1}^{p} \mathrm{E}[U_{jkl}(Z(t))] \right) \max_{1 \leqslant j,k,l \leqslant p} n\bar{\mathrm{E}}[|\dot{Z}_{ij}(t) Z_{ik}(t) Z_{il}(t)|]dt \\
&\lesssim_{(7)} (G_3 + G_2\beta + G_1\beta^2) \int_0^1 \max_{1 \leqslant j,k,l \leqslant p} n\bar{\mathrm{E}}[|\dot{Z}_{ij}(t) Z_{ik}(t) Z_{il}(t)|]dt,
\end{aligned}
$$

where (4) follows from Lemma 42, (5) from definition of $\bar{\mathrm{E}}$, (6) from a trivial inequality, (7) from Lemma 41. We have to bound the integral on the last line. Let $\omega(t) = 1/(\sqrt{t} \wedge \sqrt{1-t})$, and observe that

$$
\begin{aligned}
&\int_0^1 \max_{1 \leqslant j,k,l \leqslant p} n\bar{\mathrm{E}}[|\dot{Z}_{ij}(t) Z_{ik}(t) Z_{il}(t)|]dt \\
&= \int_0^1 \omega(t) \max_{1 \leqslant j,k,l \leqslant p} n\bar{\mathrm{E}}[|(\dot{Z}_{ij}(t)/\omega(t)) Z_{ik}(t) Z_{il}(t)|]dt \\
&\leqslant n \int_0^1 \omega(t) \max_{1 \leqslant j,k,l \leqslant p} \left(\bar{\mathrm{E}}[|\dot{Z}_{ij}(t)/\omega(t)|^3] \bar{\mathrm{E}}[|Z_{ik}(t)|^3] \bar{\mathrm{E}}[|Z_{il}(t)|^3] \right)^{1/3} dt,
\end{aligned}
$$

where the last inequality is by Hölder. The last term is further bounded as

$$
\begin{aligned}
&\leqslant_{(1)} n^{-1/2} \left\{ \int_0^1 \omega(t)dt \right\} \max_{1 \leqslant j \leqslant p} \bar{\mathrm{E}}[(|\tilde{x}_{ij}| + |\tilde{y}_{ij}|)^3] \\
&\lesssim_{(2)} n^{-1/2} \max_{1 \leqslant j \leqslant p} [(\bar{\mathrm{E}}[|\tilde{x}_{ij}|^3])^{1/3} + (\bar{\mathrm{E}}[|\tilde{y}_{ij}|^3])^{1/3}]^3 \\
&\lesssim_{(3)} n^{-1/2} \max_{1 \leqslant j \leqslant p} [(\bar{\mathrm{E}}[|x_{ij}|^3])^{1/3} + (\bar{\mathrm{E}}[|y_{ij}|^3])^{1/3}]^3 \\
&\lesssim_{(4)} n^{-1/2} \max_{1 \leqslant j \leqslant p} \bar{\mathrm{E}}[|x_{ij}|^3],
\end{aligned}
$$

where (1) follows from the fact that: $|\dot{Z}_{ij}(t)/\omega(t)| \leqslant (|\tilde{x}_{ij}| + |\tilde{y}_{ij}|)/\sqrt{n}$, $|Z_{im}(t)| \leqslant (|\tilde{x}_{im}| + |\tilde{y}_{im}|)/\sqrt{n}$, and the product of terms $\bar{\mathrm{E}}[(|\tilde{x}_{ij}| + |\tilde{y}_{ij}|)^3]^{1/3}$, $\bar{\mathrm{E}}[(|\tilde{x}_{ik}| + |\tilde{y}_{ik}|)^3]^{1/3}$ and $\bar{\mathrm{E}}[(|\tilde{x}_{il}| + |\tilde{y}_{il}|)^3]^{1/3}$ is trivially bounded by $\max_{1 \leqslant j \leqslant p} \bar{\mathrm{E}}[(|\tilde{x}_{ij}| + |\tilde{y}_{ij}|)^3]$; (2) follows from $\int_0^1 \omega(t)dt \lesssim 1$, (3) from Lemma 43 (a), and (4) from the normality of y_{ij} with

$\mathrm{E}[y_{ij}^2] = \mathrm{E}[x_{ij}^2]$, so that $\mathrm{E}[|y_{ij}|^3] \lesssim (\mathrm{E}[y_{ij}^2])^{3/2} = (\mathrm{E}[|x_{ij}^2|])^{3/2} \leqslant \mathrm{E}[|x_{ij}|^3]$. This completes the overall proof. □

3.6.4 Proof of Corollary 4

See Supplemental Appendix 3.10.2. □

3.6.5 Proof of Lemma 33

Since $\bar{\mathrm{E}}[x_{ij}^2] \geqslant c_1$ by assumption, we have $1\{|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\} \leqslant 1\{|x_{ij}| > c_1^{1/2}u\}$. By Markov's inequality and the condition of the lemma, we have

$$\begin{aligned}\mathrm{P}\left(|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}, \text{ for some } (i,j)\right) &\leqslant \textstyle\sum_{i=1}^n \mathrm{P}\left(\max_{1\leqslant j\leqslant p} |x_{ij}| > c_1^{1/2}u\right) \\ &\leqslant \textstyle\sum_{i=1}^n \mathrm{P}\left(h(\max_{1\leqslant j\leqslant p} |x_{ij}|/D) > h(c_1^{1/2}u/D)\right) \leqslant n/h(c_1^{1/2}u/D).\end{aligned}$$

This implies $u_x(\gamma) \leqslant c_1^{-1/2} D h^{-1}(n/\gamma)$. For $u_y(\gamma)$, by $y_{ij} \sim N(0, \mathrm{E}[x_{ij}^2])$ with $\mathrm{E}[x_{ij}^2] \leqslant B^2$, we have $\mathrm{E}[\exp(y_{ij}^2/(4B^2))] \lesssim 1$. Hence

$$\begin{aligned}\mathrm{P}\left(|y_{ij}| > u(\bar{\mathrm{E}}[y_{ij}^2])^{1/2}, \text{ for some } (i,j)\right) &\leqslant \textstyle\sum_{i=1}^n \sum_{j=1}^p \mathrm{P}(|y_{ij}| > c_1^{1/2}u) \\ &\leqslant \textstyle\sum_{i=1}^n \sum_{j=1}^p \mathrm{P}(|y_{ij}|/(2B) > c_1^{1/2}u/(2B)) \lesssim np\exp(-c_1u^2/(4B^2)).\end{aligned}$$

Therefore, $u_y(\gamma) \leqslant CB\sqrt{\log(pn/\gamma)}$ where $C > 0$ depends only on c_1. □

3.6.6 Proof of Corollary 5

Case (i) follows directly from Corollary 3. Hence we only consider cases (ii)-(v).

Step 1. In this step, in each case of conditions (E.2)-(E.5), we shall compute the following bounds on moments M_3 and M_4 and parameters B and D in Lemma 33 with specific choice of h:

(E.2) $B \vee M_3^3 \vee M_4^2 \leqslant C$, $D \leqslant C \log p$, $h(v) = e^v - 1$;

(E.3) $B = B_n$, $D \leqslant CB_n$, $M_3^3 \vee M_4^2 \leqslant CB_n$, $h(v) = e^v - 1$;

(E.4) $B \vee M_3^3 \vee M_4^2 \leqslant CB_n$, $D \leqslant CB_n \log p$, $h(v) = e^v - 1$;

(E.5) $B \vee D \vee M_3^3 \vee M_4^2 \leqslant CB_n$, $h(v) = v^4$.

Here $C > 0$ is a (sufficiently large) constant that depends only on c_1 and C_1. The bounds on B, M_3 and M_4 follow from elementary computations using Hölder's inequality. The bounds on D follow from an elementary application of Lemma 2.2.2 in [111]. For brevity, we omit the detail.

Step 2. In either case of (ii)-(v), there are sufficiently small constants $c_3 > 0$ and $c_4 > 0$, and a sufficiently large constant $C_2 > 0$, depending only on c_1, c_2, C_1 such that, with $\ell_n := \log(pn^{1+c_3})$,

$$n^{-1/2}\ell_n^{3/2}\max\{B\ell_n^{1/2}, Dh^{-1}(n^{1+c_3})\} \leqslant C_2 n^{-c_4},$$
$$n^{-1/8}(M_3^{3/4} \vee M_4^{1/2})\ell_n^{7/8} \leqslant C_2 n^{-c_4}.$$

Hence taking $\gamma = n^{-c_3}$, we conclude from Corollary 4 and Lemma 33 that $\rho \leqslant Cn^{-\min\{c_3,c_4\}}$ where $C > 0$ depends only on c_1, c_2, C_1. □

3.7 Appendix C. Proofs for Section 3.3

3.7.1 Proof of Lemma 35

Recall that $\Delta = \max_{1\leqslant j,k\leqslant p}|\mathbb{E}_n[x_{ij}x_{ik}] - \bar{\mathrm{E}}[x_{ij}x_{ik}]|$. By Lemma 34, on the event $\{(x_i)_{i=1}^n : \Delta \leqslant \vartheta\}$, we have $|\mathrm{P}(Z_0 \leqslant t) - \mathrm{P}_e(W_0 \leqslant t)| \leqslant \pi(\vartheta)$ for all $t \in \mathbb{R}$, and so on this event

$$\mathrm{P}_e(W_0 \leqslant c_{Z_0}(\alpha + \pi(\vartheta))) \geqslant \mathrm{P}(Z_0 \leqslant c_{Z_0}(\alpha + \pi(\vartheta))) - \pi(\vartheta) \geqslant \alpha + \pi(\vartheta) - \pi(\vartheta) = \alpha,$$

implying the first claim. The second claim follows similarly. □

3.7.2 Proof of Lemma 36

By equation (3.15), the probability of the event $\{(x_i)_{i=1}^n : \mathrm{P}_e(|W - W_0| > \zeta_1) \leqslant \zeta_2\}$ is at least $1 - \zeta_2$. On this event,

$$\mathrm{P}_e(W \leqslant c_{W_0}(\alpha + \zeta_2) + \zeta_1) \geqslant \mathrm{P}_e(W_0 \leqslant c_{W_0}(\alpha + \zeta_2)) - \zeta_2 \geqslant \alpha + \zeta_2 - \zeta_2 = \alpha,$$

implying that $\mathrm{P}(c_W(\alpha) \leqslant c_{W_0}(\alpha + \zeta_2) + \zeta_1) \geqslant 1 - \zeta_2$. The second claim of the lemma follows similarly. □

3.7.3 Proof of Theorem 15

For $\vartheta > 0$, let $\pi(\vartheta) := C_2\vartheta^{1/3}(1 \vee \log(p/\vartheta))^{2/3}$ as defined in Lemma 35. Then

$$\begin{aligned} \mathrm{P}(T_0 \leqslant c_{W_0}(\alpha)) &\leqslant_{(1)} \mathrm{P}(T_0 \leqslant c_{Z_0}(\alpha + \pi(\vartheta))) + \mathrm{P}(\Delta > \vartheta) \\ &\leqslant_{(2)} \alpha + \pi(\vartheta) + \mathrm{P}(\Delta > \vartheta) + \rho, \end{aligned}$$

where (1) follows from Lemma 35 and (2) follows from definition of ρ and the fact that Z_0 has no point masses. The upper bound is proven. The lower bound follows similarly. □

3.7.4 Proof of Theorem 16

For $\vartheta > 0$, let $\pi(\vartheta) := C_2\vartheta^{1/3}(1 \vee \log(p/\vartheta))^{2/3}$ with $C_2 > 0$ as in Lemma 35. Then

$$\begin{aligned} &\mathrm{P}(T \leqslant c_W(\alpha)) \leqslant_{(1)} \mathrm{P}(T_0 \leqslant c_W(\alpha) + \zeta_1) + \zeta_2 \\ &\leqslant_{(2)} \mathrm{P}(T_0 \leqslant c_{W_0}(\alpha + \zeta_2) + 2\zeta_1) + 2\zeta_2 \\ &\leqslant_{(3)} \mathrm{P}(T_0 \leqslant c_{Z_0}(\alpha + \zeta_2 + \pi(\vartheta)) + 2\zeta_1) + 2\zeta_2 + \mathrm{P}(\Delta > \vartheta) \\ &\leqslant_{(4)} \mathrm{P}(Z_0 \leqslant c_{Z_0}(\alpha + \zeta_2 + \pi(\vartheta)) + 2\zeta_1) + \rho + 2\zeta_2 + \mathrm{P}(\Delta > \vartheta) \\ &\leqslant_{(5)} \mathrm{P}(Z_0 \leqslant c_{Z_0}(\alpha + \zeta_2 + \pi(\vartheta))) + C_3\zeta_1\sqrt{1 \vee \log(p/\zeta_1)} + \rho + 2\zeta_2 + \mathrm{P}(\Delta > \vartheta) \\ &\leqslant_{(6)} \alpha + \zeta_2 + \pi(\vartheta) + C_3\zeta_1\sqrt{1 \vee \log(p/\zeta_1)} + 2\zeta_2 + \mathrm{P}(\Delta > \vartheta) + \rho \end{aligned}$$

where $C_3 > 0$ depends on c_1 and C_1 only and where (1) follows from equation (3.14), (2) from Lemma 36, (3) from Lemma 35, (4) from the definition of ρ, and (5) follows from Lemma 32 on anti-concentration, and (6) by the fact that Z_0 has no point masses. This gives the upper bound. The lower bound follows similarly. □

3.7.5 Proof of Corollary 6

The proof of this corollary relies on:

Lemma 44. *Recall conditions (E.2)-(E.5) in Section 3.2.2. Then*

$$\mathrm{E}[\Delta] \leqslant C \times \begin{cases} \sqrt{\frac{\log p}{n}} \vee \frac{(\log(pn))^2(\log p)}{n}, & \textit{under (E.2)}, \\ \sqrt{\frac{B_n^2 \log p}{n}} \vee \frac{B_n^2(\log p)}{n}, & \textit{under (E.3)}, \\ \sqrt{\frac{B_n^2 \log p}{n}} \vee \frac{B_n^2(\log(pn))^2(\log p)}{n}, & \textit{under (E.4)}, \\ \sqrt{\frac{B_n^2 \log p}{n}} \vee \frac{B_n^2(\log p)}{\sqrt{n}}, & \textit{under (E.5)}, \end{cases}$$

where $C > 0$ depends only on c_1 and C_1 that appear in (E.2)-(E.5).

Proof. By Lemma 37 and Hölder's inequality, we have

$$\mathrm{E}[\Delta] \lesssim M_4^2\sqrt{(\log p)/n} + (\mathrm{E}[\max_{i,j} |x_{ij}|^4])^{1/2}(\log p)/n.$$

The conclusion of the lemma follows from elementary calculations with help of Lemma 2.2.2 in [111]. □

Proof of Corollary 6. We make use of Theorem 16. Let $c > 0$ and $C > 0$ denote generic constants depending only on c_1, c_2, C_1, and their values may change from place to place. By Corollary 5, in either case of (i)-(iv), $\rho \leqslant Cn^{-c}$. Moreover, $\zeta_1\sqrt{\log p} \leqslant C_1 n^{-c_2}$ implies that $\zeta_1 \leqslant C_1 n^{-c_2}$ (recall $p \geqslant 3$), and hence $\zeta_1\sqrt{\log(p/\zeta_1)} \leqslant Cn^{-c}$. Also, $\zeta_2 \leqslant Cn^{-c}$ by assumption.

Let $\vartheta = \vartheta_n := (\mathrm{E}[\Delta])^{1/2}/\log p$. By Lemma 44, $\mathrm{E}[\Delta](\log p)^2 \leqslant Cn^{-c}$. Therefore, $\pi(\vartheta) \leqslant Cn^{-c}$ (with possibly different $c, C > 0$). In addition, by Markov's inequality,

$\mathrm{P}(\Delta > \vartheta) \leqslant \mathrm{E}[\Delta]/\vartheta \leqslant Cn^{-c}$. Hence, by Theorem 16, we have $\sup_{\alpha\in(0,1)} |\mathrm{P}(T \leqslant c_W(\alpha)) - \alpha| \leqslant Cn^{-c}$. □

3.8 Appendix D. Proofs for Section 3.4

3.8.1 Proof of Theorem 17

The proof proceeds in three steps. In the proof $(\widehat{\beta}, \lambda)$ denotes $(\widehat{\beta}^{(k)}, \lambda^{(k)})$ with k either 0 or 1.

Step 1. Here we show that there exist some constants $c > 0$ and $C > 0$ (depending only c_1, C_1 and σ^2) such that for either $k \in \{0, 1\}$,

$$\mathrm{P}(T_0 \leqslant \lambda^{(k)}) \geqslant 1 - \alpha - \nu_n, \tag{3.21}$$

with $\nu_n = Cn^{-c}$. We first note that $T_0 = \sqrt{n} \max_{1\leqslant k\leqslant 2p} \mathbb{E}_n[\tilde{z}_{ik}\varepsilon_i]$, where $\tilde{z}_i = (z_i', -z_i')'$. Application of Corollary 5-(v) gives

$$|\mathrm{P}(T_0 \leqslant \lambda) - \mathrm{P}(Z_0 \leqslant \lambda)| \leqslant Cn^{-c},$$

where $c > 0$ and $C > 0$ are constants depending only on c_1, C_1 and σ^2. Since $\lambda \geqslant c_{Z_0}(1-\alpha)$, the claim follows. Indeed, $\lambda^{(1)} = c_{Z_0}(1-\alpha)$, and $\lambda^{(1)} \leqslant \lambda^{(0)} = c_0(1-\alpha) := \sigma\Phi^{-1}(1-\alpha/(2p))$, since by the union bound $\mathrm{P}(Z_0 \geqslant c_0(1-\alpha)) \leqslant 2p\mathrm{P}(\sigma N(0,1) \geqslant c_0(1-\alpha)) = \alpha$.

Step 2. We claim that with probability $\geqslant 1 - \alpha - \nu_n$, $\widehat{\delta} = \widehat{\beta} - \beta$ obeys:

$$\sqrt{n} \max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}(z_i'\widehat{\delta})]| \leqslant 2\lambda.$$

Indeed, by definition of $\widehat{\beta}$, $\sqrt{n}\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}(y_i - z_i'\widehat{\beta})]| \leqslant \lambda$, which by the triangle inequality implies $\sqrt{n}\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}(z_i'\widehat{\delta})]| \leqslant T_0 + \lambda$. The claim follows from Step 1.

Step 3. By Step 1, with probability $\geqslant 1 - \alpha - \nu_n$, the true value β obeys the constraint in optimization problem (3.16), in which case by definition of $\widehat{\beta}$, $\|\widehat{\beta}\|_{\ell_1} \leqslant$

$\|\beta\|_{\ell_1}$. Therefore, with the same probability, $\widehat{\delta} \in \mathcal{R}(\beta) = \{\delta \in \mathbb{R}^d : \|\beta+\delta\|_{\ell_1} \leqslant \|\beta\|_{\ell_1}\}$. By definition of $\kappa_I(\beta)$ we have that

$$\kappa_I(\beta)\|\widehat{\delta}\|_I \leqslant \max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}(z_i'\widehat{\delta})]|.$$

Combining this inequality with Step 2 gives the claim of the theorem. □

3.8.2 Proof of Theorem 18

The proof has four steps. In the proof, we let $\varrho_n = Cn^{-c}$ for sufficiently small $c > 0$ and sufficiently large $C > 0$ depending only on $c_1, C_1, \underline{\sigma}^2, \sigma^2$, where c and C (and hence ϱ_n) may change from place to place.

Step 0. The same argument as in the previous proof applies to $\widehat{\beta}^{(0)}$ with $\lambda = \lambda^{(0)} := c_0(1-1/n)$, where now σ^2 is the upper bound on $\mathrm{E}[\varepsilon_i^2]$. Thus, we conclude that with probability at least $1-\varrho_n$,

$$\|\widehat{\beta}^{(0)} - \beta\|_{\mathrm{pr}} \leqslant \frac{2c_0(1-1/n)}{\sqrt{n}\kappa_{\mathrm{pr}}(\beta)}.$$

Step 1. We claim that with probability at least $1-\varrho_n$,

$$\max_{1\leqslant j\leqslant p}\left(\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2]\right)^{1/2} \leqslant B_n\frac{2c_0(1-1/n)}{\sqrt{n}\kappa_{\mathrm{pr}}(\beta)} =: \iota_n.$$

Application of Hölder's inequality and identity $\varepsilon_i - \widehat{\varepsilon}_i = z_i'(\widehat{\beta}^{(0)} - \beta)$ gives

$$\max_{1\leqslant j\leqslant p}\left(\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2]\right)^{1/2} \leqslant B_n(\mathbb{E}_n[z_i'(\widehat{\beta}^{(0)}-\beta)]^2)^{1/2} \leqslant B_n\|\widehat{\beta}^{(0)}-\beta\|_{\mathrm{pr}}.$$

The claim follows from Step 0.

Step 2. In this step, we apply Corollary 6-(iv) to

$$T = T_0 = \sqrt{n}\max_{1\leqslant j\leqslant 2p}\mathbb{E}_n[\tilde{z}_{ij}\varepsilon_i],\ W = \sqrt{n}\max_{1\leqslant j\leqslant 2p}\mathbb{E}_n[\tilde{z}_{ij}\widehat{\varepsilon}_i e_i], \text{ and}$$
$$W_0 = \sqrt{n}\max_{1\leqslant j\leqslant 2p}\mathbb{E}_n[\tilde{z}_{ij}\varepsilon_i e_i],$$

where $\tilde{z}_i = (z_i', -z_i')'$, to conclude that uniformly in $\alpha \in (0,1)$

$$\mathrm{P}(T_0 \leqslant c_W(1-\alpha)) \geqslant 1 - \alpha - \varrho_n. \tag{3.22}$$

To show applicability of Corollary 6-(iv), we note that for any $\zeta_1 > 0$,

$$\begin{aligned} \mathrm{P}_e(|W - W_0| > \zeta_1) \leqslant \mathrm{E}_e[|W - W_0|]/\zeta_1 \leqslant \sqrt{n}\mathrm{E}_e \left[\max_{1 \leqslant j \leqslant p} |\mathbb{E}_n[z_{ij}(\widehat{\varepsilon}_i - \varepsilon_i)e_i]| \right] /\zeta_1 \\ \lesssim \sqrt{\log p} \max_{1 \leqslant j \leqslant p} (\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2])^{1/2}/\zeta_1, \end{aligned}$$

where the third inequality is due to Pisier's inequality. The last quantity is bounded by $(\iota_n^2 \log p)^{1/2}/\zeta_1$ with probability $\geqslant 1 - \varrho_n$ by Step 1.

Since $\iota_n \log p \leqslant C_1 n^{-c_1}$ by assumption (vi) of the theorem, we can take ζ_1 in such a way that $\zeta_1(\log p)^{1/2} \leqslant \varrho_n$ and $(\iota_n^2 \log p)^{1/2}/\zeta_1 \leqslant \varrho_n$. Then all the conditions of Corollary 6-(iv) with so defined ζ_1 and $\zeta_2 = \varrho_n \vee ((\iota_n^2 \log p)^{1/2}/\zeta_1)$ are satisfied, and hence application of the corollary gives that uniformly in $\alpha \in (0,1)$,

$$|\mathrm{P}(T_0 \leqslant c_W(1-\alpha)) - 1 - \alpha| \leqslant \varrho_n, \tag{3.23}$$

which implies the claim of this step.

Step 3. In this step we claim that with probability at least $1 - \varrho_n$,

$$c_W(1-\alpha) \leqslant c_{Z_0}(1 - \alpha + 2\varrho_n).$$

Combining Step 2 and Lemma 36 gives that with probability at least $1 - \zeta_2$, $c_W(1 - \alpha) \leqslant c_{W_0}(1 - \alpha + \zeta_2) + \zeta_1$, where ζ_1 and ζ_2 are chosen as in Step 2. In addition, Lemma 35 shows that $c_{W_0}(1 - \alpha + \zeta_2) \leqslant c_{Z_0}(1 - \alpha + \varrho_n)$. Finally, Lemma 32 yields $c_{Z_0}(1 - \alpha + \varrho_n) + \zeta_1 \leqslant c_{Z_0}(1 - \alpha + 2\varrho_n)$. Combining these bounds gives the claim of this step.

Step 4. Given (3.22), the rest of the proof is identical to Steps 2-3 in the proof of Theorem 17 with $\lambda = c_W(1-\alpha)$. The result follows for $\nu_n = 2\varrho_n$. □

3.9 Appendix E. A note on relation between Slepian and Stein type methods for normal approximations

To keep the notation simple, consider a random vector X in $\mathbb{R}^p$ and a standard normal vector Z in $\mathbb{R}^p$. We are interested in bounding

$$\mathrm{E}[g(X)] - \mathrm{E}[g(Z)],$$

over some collection of test functions $g \in \mathcal{G}$. Without loss of generality, suppose that Z and X are independent.

Consider Stein's partial differential equation:

$$g(x) - \mathrm{E}[g(Z)] = \Delta h(x) - x'\nabla h(x).$$

It is well known, e.g. [52] and [29], that an explicit solution for h in this equation is given by

$$h(x) := -\int_0^1 \frac{1}{2t}\left[\mathrm{E}[g(\sqrt{t}x + \sqrt{1-t}Z)] - \mathrm{E}[g(Z)]\right] dt,$$

so that

$$\mathrm{E}[g(X)] - \mathrm{E}[g(Z)] = \mathrm{E}[\Delta h(X) - X'\nabla h(X)].$$

The Stein type method for normal approximation bounds the right side for $g \in \mathcal{G}$.

Next, let us consider the Slepian smart path interpolation:

$$Z(t) = \sqrt{t}X + \sqrt{1-t}Z.$$

Then we have

$$\mathrm{E}[g(X)] - \mathrm{E}[g(Z)] = \mathrm{E}\left[\int_0^1 \frac{1}{2}\nabla g(Z(t))'\left(\frac{X}{\sqrt{t}} - \frac{Z}{\sqrt{1-t}}\right)\right] dt.$$

The Slepian type method, as used in our paper, bounds the right side for $g \in \mathcal{G}$.

Elementary calculations and integration by parts yield the following observation.

Lemma 45. *Suppose that $g : \mathbb{R}^p \to \mathbb{R}$ is a C^2-function with uniformly bounded derivatives up to order two. Then*

$$I := \mathrm{E}\left[\int_0^1 \frac{1}{2}\nabla g(Z(t))'\left(\frac{X}{\sqrt{t}}\right)\right] = -\mathrm{E}[X'\nabla h(X)]$$

and

$$II := \mathrm{E}\left[\int_0^1 \frac{1}{2}\nabla g(Z(t))'\left(\frac{Z}{\sqrt{1-t}}\right)\right] = -\mathrm{E}[\triangle h(X)].$$

Hence the Slepian and Stein methods both show that difference between I and II is small or approaches zero under suitable conditions on X; therefore, they are very similar in spirit, if not identical. The details of treating terms may be different from application to application.

Proof of Lemma 45. By definition of h, we have

$$-\mathrm{E}[X'\nabla h(X)] = \mathrm{E}\left[X'\int_0^1 \frac{1}{2t}\nabla g(Z(t))\sqrt{t}dt\right] = \mathrm{E}\left[\int_0^1 \nabla g(Z(t))'\frac{X}{2\sqrt{t}}dt\right].$$

On the other hand, by definition of h and Stein's identity (Lemma 46),

$$-\mathrm{E}[\triangle h(X)] = \mathrm{E}\left[\frac{1}{2}\int_0^1 \triangle g(Z(t))dt\right] = \mathrm{E}\left[\frac{1}{2}\int_0^1 \nabla g(Z(t))'\left(\frac{Z}{\sqrt{1-t}}\right)dt\right].$$

This completes the proof. □

Lemma 46 (Stein's identity). *Let $W = (W_1, \ldots, W_p)^T$ be a centered Gaussian random vector in $\mathbb{R}^p$. Let $f : \mathbb{R}^p \to \mathbb{R}$ be a C^1-function such that $\mathrm{E}[|\partial_j f(W)|] < \infty$ for all $1 \leqslant j \leqslant p$. Then for every $1 \leqslant j \leqslant p$,*

$$\mathrm{E}[W_j f(W)] = \sum_{k=1}^p \mathrm{E}[W_j W_k]\mathrm{E}[\partial_k f(W)].$$

Proof of Lemma 46. See Section A.6 of [108], and also [107]. □

3.10 Appendix F. Additional proofs

3.10.1 Proof of Lemma 43

Claim (a). Define $I_{ij} = 1\{|x_{ij}| \leqslant u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\}$, and observe that

$$\begin{aligned}(\bar{\mathrm{E}}[|\tilde{x}_{ij}|^q])^{1/q} &\leqslant (\bar{\mathrm{E}}[|x_{ij}I_{ij}|^q])^{1/q} + (\mathbb{E}_n[|\mathrm{E}[x_{ij}I_{ij}]|^q])^{1/q} \\ &\leqslant (\bar{\mathrm{E}}[|x_{ij}I_{ij}|^q])^{1/q} + (\bar{\mathrm{E}}[|x_{ij}I_{ij}|^q])^{1/q} \leqslant 2(\bar{\mathrm{E}}[|x_{ij}|^q])^{1/q}.\end{aligned}$$

Claim (b). Observe that

$$\begin{aligned}\bar{\mathrm{E}}[|\tilde{x}_{ij}\tilde{x}_{ik} - x_{ij}x_{ik}|] &\leqslant \bar{\mathrm{E}}[|(\tilde{x}_{ij} - x_{ij})\tilde{x}_{ik}|] + \bar{\mathrm{E}}[|x_{ij}(\tilde{x}_{ik} - x_{ik})|] \\ &\leqslant \sqrt{\bar{\mathrm{E}}[(\tilde{x}_{ij} - x_{ij})^2]}\sqrt{\bar{\mathrm{E}}[\tilde{x}_{ik}^2]} + \sqrt{\bar{\mathrm{E}}[(\tilde{x}_{ik} - x_{ik})^2]}\sqrt{\bar{\mathrm{E}}[x_{ij}^2]} \\ &\leqslant 2\varphi(u)\sqrt{\bar{\mathrm{E}}[x_{ij}^2]}\sqrt{\bar{\mathrm{E}}[x_{ik}^2]} + \varphi(u)\sqrt{\bar{\mathrm{E}}[x_{ik}^2]}\sqrt{\bar{\mathrm{E}}[x_{ij}^2]} \\ &\leqslant (3/2)\varphi(u)(\bar{\mathrm{E}}[x_{ij}^2] + \bar{\mathrm{E}}[x_{ik}^2]),\end{aligned}$$

where the first inequality follows from the triangle inequality, the second from the Cauchy-Schwarz inequality, the third from the definition of $\varphi(u)$ together with claim (a), and the last from inequality $|ab| \leqslant (a^2 + b^2)/2$.

Claim (c). This follows from the Cauchy-Schwarz inequality.

Claim (d). We shall use the following lemma.

Lemma 47 (Tail Bounds for Self-Normalized Sums). *Let $\xi_1, \ldots, \xi_n$ be independent real-valued random variables such that $\mathrm{E}[\xi_i] = 0$ and $\mathrm{E}[\xi_i^2] < \infty$ for all $1 \leqslant i \leqslant n$. Let $S_n = \sum_{i=1}^n \xi_i$. Then for every $x > 0$,*

$$\mathrm{P}(|S_n| > x(4B_n + V_n)) \leqslant 4\exp(-x^2/2),$$

where $B_n^2 = \sum_{i=1}^n \mathrm{E}[\xi_i^2]$ and $V_n^2 = \sum_{i=1}^n \xi_i^2$.

Proof of Lemma 47. See [39], Theorem 2.16. □

Define

$$\Lambda_j := 4\sqrt{\bar{\mathrm{E}}[(x_{ij} - \tilde{x}_{ij})^2]} + \sqrt{\mathbb{E}_n[(x_{ij} - \tilde{x}_{ij})^2]}.$$

Then by Lemma 47 and the union bound, with probability at least $1 - 4\gamma$,

$$|X_j - \tilde{X}_j| \leqslant \Lambda_j \sqrt{2\log(p/\gamma)}, \text{ for all } 1 \leqslant j \leqslant p.$$

By claim (c), for $u \geqslant u(\gamma)$, with probability at least $1 - \gamma$, for all $1 \leqslant j \leqslant p$,

$$\Lambda_j = 4\sqrt{\bar{\mathrm{E}}[(x_{ij} - \tilde{x}_{ij})^2]} + \sqrt{\mathbb{E}_n[(\mathrm{E}[x_{ij} - \tilde{x}_{ij}])^2]} \leqslant 5\sqrt{\bar{\mathrm{E}}[x_{ij}^2]}\varphi(u).$$

The last two assertions imply claim (d). □

3.10.2 Proof of Corollary 4

Since M_2 is bounded from below and above by positive constants, we may normalize $M_2 = 1$, without loss of generality. In this proof, let $C > 0$ denote a generic constant depending only on c_1 and C_1, and its value may change from place to place.

For given $\gamma \in (0, 1)$, denote $\ell_n := \log(pn/\gamma) \geqslant 1$ and let

$$u_1 := n^{3/8}\ell_n^{-5/8}M_3^{3/4} \text{ and } u_2 := n^{3/8}\ell_n^{-5/8}M_4^{1/2}.$$

Define $u := u(\gamma) \vee u_1 \vee u_2$ and $\beta := \sqrt{n}/(2\sqrt{2}u)$. Then $u \geqslant u(\gamma)$ and the choice of β trivially obeys $2\sqrt{2}u\beta \leqslant \sqrt{n}$. So, by Theorem 14 and using the argument as that in the proof of Corollary 3, for every $\psi > 0$, we have for any $\bar{\varphi}(u) \geqslant \varphi(u)$,

$$\begin{aligned}\rho \leqslant C\big[&n^{-1/2}(\psi^3 + \psi^2\beta + \psi\beta^2)M_3^3 + (\psi^2 + \psi\beta)\bar{\varphi}(u) \\ &+\psi\bar{\varphi}(u)\sqrt{\log(p/\gamma)} + (\beta^{-1}\log p + \psi^{-1})\sqrt{1 \vee \log(p\psi)}\big].\end{aligned} \qquad (3.24)$$

Step 1. We claim that we can take $\bar{\varphi}(u) := CM_4^2/u$ for all $u > 0$. Since

$\bar{\mathrm{E}}[x_{ij}^2] \geqslant c_1$, we have $1\{|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\} \leqslant 1\{|x_{ij}| > c_1^{1/2}u\}$. Hence

$$\begin{aligned}\bar{\mathrm{E}}[x_{ij}^2 1\{|x_{ij}| > u(\bar{\mathrm{E}}[x_{ij}^2])^{1/2}\}] &\leqslant \bar{\mathrm{E}}[x_{ij}^2 1\{|x_{ij}| > c_1^{1/2}u\}] \\ &\leqslant \bar{\mathrm{E}}[x_{ij}^4 1\{|x_{ij}| > c_1^{1/2}u\}]/(c_1u^2) \leqslant \bar{\mathrm{E}}[x_{ij}^4]/(c_1u^2) \leqslant M_4^4/(c_1u^2).\end{aligned}$$

This implies $\varphi_x(u) \leqslant CM_4^2/u$. For $\varphi_y(u)$, note that

$$\bar{\mathrm{E}}[y_{ij}^4] = \mathbb{E}_n[\mathrm{E}[y_{ij}^4]] = 3\mathbb{E}_n[(\mathrm{E}[y_{ij}^2])^2] = 3\mathbb{E}_n[(\mathrm{E}[x_{ij}^2])^2] \leqslant 3\mathbb{E}_n[\mathrm{E}[x_{ij}^4]] = \bar{\mathrm{E}}[x_{ij}^4],$$

and hence $\varphi_y(u) \leqslant CM_4^2/u$ as well. This implies the claim of this step.

Step 2. We shall bound the right side of (3.24) by suitably choosing ψ depending on the range of u. In order to set up this choice we define $u^\star$ by the following equation:

$$\bar{\varphi}(u^\star)n^{3/8}/(M_3^3\ell_n^{5/6})^{3/4} = 1.$$

We then take

$$\psi = \psi(u) := \begin{cases} n^{1/8}\ell_n^{-3/8}M_3^{-3/4} & \text{if } u \geqslant u^\star, \\ \ell_n^{-1/6}(\bar{\varphi}(u))^{-1/3} & \text{if } u < u^\star. \end{cases} \tag{3.25}$$

We note that for $u < u^\star$,

$$\psi(u) \leqslant \psi(u^\star) = n^{1/8}\ell_n^{-3/8}M_3^{-3/4}.$$

That is, when $u < u^\star$ the smoothing parameter ψ is smaller than when $u \geqslant u^\star$.

Using these choices of parameters β and ψ and elementary calculations (which will be done in Step 3 below), we conclude from (3.24) that whether $u < u^\star$ or $u \geqslant u^\star$,

$$\rho \leqslant C(n^{-1/2}u\ell_n^{3/2} + \gamma).$$

The bound in the corollary follows from this inequality.

Step 3. (Computation of the bound on ρ). Note that since $\rho \leqslant 1$, we only had to consider the case where $n^{-1/2}u\ell_n^{3/2} \leqslant 1$ since otherwise the inequality is trivial by taking, say, $C = 1$. Since $u_1 = n^{3/8}M_3^{3/4}/\ell_n^{5/8}$ and $u_2 = n^{3/8}M_4^{1/2}/\ell_n^{5/8}$, we have

$$\begin{aligned}
(\bar{\varphi}(u^\star))^{4/3} &= n^{-1/2}\ell_n^{5/6}M_3^3, \\
\bar{\varphi}(u_1) &\leqslant Cn^{-3/8}\ell_n^{5/8}M_4^2/M_3^{3/4}, \\
\bar{\varphi}(u_2) &\leqslant Cn^{-3/8}\ell_n^{5/8}M_4^{3/2}.
\end{aligned}$$

Also note that $\psi \leqslant n^{1/8}$, and so $1 \vee \log(p\psi) \lesssim \log(pn) \leqslant \ell_n$. Therefore,

$$\beta^{-1}\log p\sqrt{1 \vee \log(p\psi)} \lesssim \beta^{-1}\ell_n^{3/2} \lesssim n^{-1/2}u\ell_n^{3/2}.$$

In addition, note that $\beta \lesssim \sqrt{n}/u \leqslant \sqrt{n}/u_1 = n^{1/8}\ell_n^{5/8}M_3^{-3/4} =: \bar{\beta}$ and $\psi \leqslant \bar{\beta}$ under either case. This implies that $(\psi^3 + \psi^2\beta + \psi\beta^2) \lesssim \psi\bar{\beta}^2$ and $(\psi^2 + \psi\beta) \leqslant \psi\bar{\beta}$. Using these inequalities, we can compute the bounds claimed above.

(a). Bounding ρ when $u \geqslant u^\star$. Then

$$\begin{aligned}
& n^{-1/2}(\psi^3 + \psi^2\beta + \psi\beta^2)M_3^3 \lesssim n^{-1/2}\psi\bar{\beta}^2M_3^3 \leqslant n^{-1/8}\ell_n^{7/8}M_3^{3/4} \leqslant n^{-1/2}u\ell_n^{3/2}; \\
& (\psi^2 + \psi\beta)\bar{\varphi}(u) \lesssim \psi\bar{\beta}\bar{\varphi}(u) \leqslant \psi\bar{\beta}\bar{\varphi}(u^\star) \leqslant n^{-1/8}\ell_n^{7/8}M_3^{3/4} \leqslant n^{-1/2}u\ell_n^{3/2}; \\
& \psi\bar{\varphi}(u)\sqrt{\log(p/\gamma)} \leqslant \psi\bar{\beta}\bar{\varphi}(u)\sqrt{\ell_n}/\bar{\beta} \leqslant \psi\bar{\beta}\bar{\varphi}(u^\star) \leqslant n^{-1/2}u\ell_n^{3/2};\ \text{and} \\
& \psi^{-1}\sqrt{\ell_n} \leqslant n^{-1/8}\ell_n^{7/8}M_3^{3/4} \leqslant n^{-1/2}u\ell_n^{3/2};
\end{aligned}$$

where we have used Step 1 and the fact that

$$\sqrt{\ell_n}/\bar{\beta} = \ell_n^{-1/2}\psi^{-1} \leqslant n^{-1/8}\ell_n^{-1/8}M_3^{3/4} \leqslant n^{-1/2}u\ell_n^{3/2} \leqslant 1.$$

The claimed bound on ρ now follows.

(b). Bounding ρ when $u < u^\star$. Since ψ is smaller than in case (a), by the calculations in Step (a)

$$n^{-1/2}(\psi^3 + \psi^2\beta + \psi\beta^2)M_3^3/\sqrt{n} \lesssim n^{-1/2}u\ell_n^{3/2}.$$

Moreover, using definition of ψ, $u > u_2$, definition of u_2, we have

$$\psi\beta\bar{\varphi}(u) \leqslant \beta\bar{\varphi}(u)^{2/3}\ell_n^{-1/6} \leqslant \beta\bar{\varphi}(u_2)^{2/3}\ell_n^{-1/6} \leqslant n^{-1}\beta u_2^2\ell_n^{5/3-1/6} \lesssim n^{-1/2}u\ell_n^{3/2};$$
$$\psi^2\bar{\varphi}(u) \leqslant \bar{\varphi}(u)^{1/3}\ell_n^{-1/3} \leqslant \bar{\varphi}(u_2)^{1/3}\ell_n^{-1/3} \leqslant n^{-1/2}u_2\sqrt{\ell_n} \leqslant n^{-1/2}u\ell_n^{3/2}.$$

Analogously and using $n^{-1/2}u\ell_n^{3/2} \leqslant 1$, we have

$$\psi\bar{\varphi}(u)\sqrt{\log(p/\gamma)} \leqslant \bar{\varphi}(u)^{2/3}\ell_n^{1/3} \leqslant \bar{\varphi}(u_2)^{2/3}\ell_n^{1/3} \leqslant nu_2^2\ell_n^2 \leqslant n^{-1/2}u\ell_n^{3/2}.$$
$$\psi^{-1}\sqrt{\ell_n} = \bar{\varphi}(u)^{1/3}\ell_n^{2/3} \leqslant n^{-1/2}u\ell_n^{3/2}.$$

This completes the proof. □

3.11 Appendix G. Application: Multiple Hypothesis Testing via the Stepdown Method

In this section, we study the problem of multiple hypothesis testing in the framework of multiple linear regressions. (Note that the problem of testing multiple means is a special case of testing multiple regressions.) We combine a general stepdown procedure described in [103] with the multiplier bootstrap developed in this paper. In contrast with [103], our results do not require weak convergence arguments, and, thus, can be applied to models with increasing numbers of both parameters and regressions. Notably, the number of regressions can be large in comparison with the sample size.

Let $(z_i, y_i)_{i=1}^n$ be a sample of independent observations where $z_i \in \mathbb{R}^p$ is a vector of non-stochastic covariates and $y_i \in \mathbb{R}^K$ is a vector of dependent random variables. For each $k = 1, \dots, K$, let $I_k \subset \{1, \dots, p\}$ be a subset of covariates used in the k-th regression. Denote by $|I_k| = p_k$ the number of covariates in the k-th regression, and let $\bar{p} = \max_{1 \leqslant k \leqslant K} p_k$. Let v_{ik} be a subvector of z_i consisting of those elements of z_i whose indices appear in I_k: $v_{ik} = (z_{ij})_{j \in I_k}$. We denote components of v_{ik} by v_{ikj}, $j = 1, \dots, p_k$. Without loss of generality, we assume that $I_k \cap I_{k'} = \varnothing$ for all $k \neq k'$

and $\sum_{1\leqslant k\leqslant K} p_k = p$.

For each $k = 1,\dots,K$, consider the linear regression model

$$y_{ik} = v_{ik}'\beta_k + \varepsilon_{ik},\ i = 1,\dots,n,$$

where $\beta_k \in \mathbb{R}^{p_k}$ is an unknown parameter of interest, and $(\varepsilon_{ik})_{i=1}^n$ is a sequence of independent zero-mean unobservable scalar random variables. We allow for triangular array asymptotics so that everything in the model, and, in particular, the number of regressions K and the dimensions of the parameters β_k and p_k, may depend on n. For brevity, however, we omit index n. We are interested in simultaneously testing the set of null hypotheses $H_{kj} : \beta_{kj} = 0$ against the alternatives $H_{kj}' : \beta_{kj} \neq 0$, $(k,j) \in \mathcal{W}_0$ for some set of pairs $\mathcal{W}_0$ where β_{kj} denotes the jth component of β_k, with the strong control of the family-wise error rate. In other words, we seek a procedure that would reject at least one true null hypothesis with probability not greater than $\alpha + o(1)$ uniformly over the set of true null hypotheses. More formally, let Ω be a set of all data generating processes, and ω be the true process. Each null hypothesis H_{kj} is equivalent to $\omega \in \Omega_{kj}$ for some subset Ω_{kj} of Ω. Let $\mathcal{W}$ denote the set of all pairs (k,j) with $k = 1,\dots,K$ and $j = 1,\dots,p_k$:

$$\mathcal{W} = \{(k,j) : k = 1,\dots,K; j = 1,\dots,p_k\}.$$

For a subset $w \subset \mathcal{W}$ let $\Omega^w = (\cap_{(k,j)\in w}\Omega_{kj}) \cap (\cap_{(k,j)\notin w}\Omega_{kj}^c)$ where $\Omega_{kj}^c = \Omega\backslash\Omega_{kj}$. The strong control of the family-wise error rate means

$$\sup_{w\subset\mathcal{W}}\sup_{\omega\in\Omega^w} \mathrm{P}\{\text{reject at least one hypothesis among } H_{kj},\ (k,j) \in w\} \leqslant \alpha + o(1). \tag{3.26}$$

This setting is clearly of interest in many empirical studies.

Our approach is based on the simultaneous analysis of t-statistics for each component β_{kj}. Let $x_{ik} = (\mathbb{E}_n[v_{ik}v_{ik}'])^{-1}v_{ik}$. Then the OLS estimator $\widehat{\beta}_k$ of β_k is given by $\widehat{\beta}_k = \mathbb{E}_n[x_{ik}y_{ik}]$. The corresponding residuals are $\widehat{\varepsilon}_{ik} = y_{ik} - v_{ik}'\widehat{\beta}$, $i = 1,\dots,n$. Since $(x_{ik})_{i=1}^n$ is non-stochastic, the covariance matrix of $\widehat{\beta}_k$ is given by $V(\widehat{\beta}_k) =$

$\mathbb{E}_n[x_{ik}x_{ik}'\sigma_{ik}^2]/n$ where $\sigma_{ik}^2 = \mathrm{E}[\varepsilon_{ik}^2]$, $i = 1, \ldots, n$. The t-statistic for testing H_{kj} against H_{kj}' is $t_{kj} := |\widehat{\beta}_{kj}|/\sqrt{\widehat{V}(\widehat{\beta}_k)_{jj}}$ where $\widehat{V}(\widehat{\beta}_k) = \mathbb{E}_n[x_{ik}x_{ik}'\widehat{\varepsilon}_{ik}^2]/n$. Also define

$$t_{kj}^0 := \frac{|\sum_{i=1}^n x_{ikj}\varepsilon_{ik}/\sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2\widehat{\varepsilon}_{ik}^2]}}.$$

Note that $t_{kj} = t_{kj}^0$ under the hypothesis H_{kj}.

The stepdown procedure of [103] is described as follows. For a subset $w \subset \mathcal{W}$, let $c_{1-\alpha,w}$ be some estimator of the $(1-\alpha)$-quantile of $\max_{(k,j)\in w} t_{kj}^0$. On the first step, let $w(1) = \mathcal{W}_0$. Reject all hypotheses H_{kj} satisfying $t_{kj} > c_{1-\alpha,w(1)}$. If no null hypothesis is rejected, then stop. If some H_{kj} are rejected, then let $w(2)$ be the set of all null hypotheses that were not rejected on the first step. On step $l \geqslant 2$, let $w(l) \subset \mathcal{W}$ be the subset of null hypotheses that were not rejected up to step l. Reject all hypotheses H_{kj}, $(k,j) \in w(l)$, satisfying $t_{kj} > c_{1-\alpha,w(l)}$. If no null hypothesis is rejected, then stop. If some H_{kj} are rejected, then let $w(l+1)$ be the subset of all null hypotheses among $(k,j) \in w(l)$ that were not rejected. Proceed in this way until the algorithm stops.

[103] proved the following result. Suppose that $c_{1-\alpha,w}$ satisfies

$$c_{1-\alpha,w'} \leqslant c_{1-\alpha,w''} \quad \text{whenever } w' \subset w'', \tag{3.27}$$

$$\sup_{w\subset\mathcal{W}}\sup_{\omega\in\Omega^w} \mathrm{P}\left(\max_{(k,j)\in w} t_{kj}^0 > c_{1-\alpha,w}\right) \leqslant \alpha + o(1), \tag{3.28}$$

then inequality (3.26) holds. Indeed, let w be the set of true null hypotheses. Suppose that the procedure rejects at least one of these hypotheses. Let l be the step when the procedure rejected a true null hypothesis for the first time, and let $H_{k_0j_0}$ be this hypothesis. Clearly, we have $w(l) \supset w$. So,

$$\max_{(k,j)\in w} t_{kj}^0 \geqslant t_{k_0j_0}^0 = t_{k_0j_0} > c_{1-\alpha,w(l)} \geqslant c_{1-\alpha,w}.$$

Combining this chain of inequalities with (3.28) yields (3.26).

To obtain suitable $c_{1-\alpha,w}$ that satisfies inequalities (3.27) and (3.28) above, we can use the multiplier bootstrap method. Let $(e_i)_{i=1}^n$ be an i.i.d. sequence of $N(0,1)$ random variables that are independent of the data. Let $c_{1-\alpha,w}$ be the conditional $(1-\alpha)$-quantile of

$$\max_{(k,j)\in w} \frac{|\sum_{i=1}^n x_{ikj}\widehat{\varepsilon}_{ik}e_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2\widehat{\varepsilon}_{ik}^2]}} \tag{3.29}$$

given $(z_i, y_i)_{i=1}^n$. To prove that so defined critical values $c_{1-\alpha,w}$ satisfy inequalities (3.27) and (3.28), we will assume the following regularity condition,

(M) There are some constants $c_1 > 0, \bar{\sigma}^2 > 0, \underline{\sigma}^2 > 0$ and a sequence $B_n \geqslant 1$ of constants such that for $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$, $1 \leqslant k \leqslant K$, $1 \leqslant l \leqslant p_k$: (i) $|z_{ij}| \leqslant B_n$; (ii) $\mathbb{E}_n[z_{ij}^2] = 1$; (iii) $\underline{\sigma}^2 \leqslant \mathrm{E}[\varepsilon_{ik}^2] \leqslant \bar{\sigma}^2$; (iv) the minimum eigenvalue of $\mathbb{E}_n[v_{ik}v_{ik}']$ is bounded from below by c_1; and (v) $\mathbb{E}_n[x_{ikl}^2] \geqslant c_1$.

Theorem 19 (Strong Control of Family-Wise Error Rate). *Let $C_1 > 0$ be some constant and suppose that assumption M is satisfied. Moreover, suppose either*

(a) $\mathrm{E}[\max_{1\leqslant k\leqslant K} \varepsilon_{ik}^4] \leqslant C_1$ *for all* $1 \leqslant i \leqslant n$, $\bar{p}^3 B_n^4(\log p)^4/n = o(1)$ *and in addition* $\bar{p}^2 B_n^4(\log(pn))^7/n = o(1)$; *or*

(b) $\mathrm{E}[\exp(|\varepsilon_{ik}|/C_1)] \leqslant 2$ *for all* $1 \leqslant i \leqslant n$, $1 \leqslant k \leqslant K$, $\bar{p}^3 B_n^2(\log p)^3/n = o(1)$ *and* $\bar{p} B_n^2(\log(pn))^7/n = o(1)$.

Then the stepdown procedure with the multiplier bootstrap critical values $c_{1-\alpha,w}$ given above satisfies (3.26).

Comment 27 (Relation to prior results). *There is a vast literature on multiple hypothesis testing. Let us consider the simple case where $K = p, p_k = 1$ for all $k = 1, \ldots, K$ and $v_{ik} = 1$, so that the k-th regression reduces to $y_{ik} = \beta_k + \varepsilon_{ik}$ (here β_k is scalar). The problem then reduces to testing multiple means (without stepdown). It is instructive to see the implication of Theorem 19 in this simple setting. Denote by t_k the t-statistic for testing $H_k : \beta_k = 0$ against $H_k' : \beta_k \neq 0$, and let $c_{1-\alpha}$ be the conditional $(1-\alpha)$-quantile of*

$$\max_{k=1,\ldots,p} \frac{|\sum_{i=1}^n \widehat{\varepsilon}_{ik}e_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[\widehat{\varepsilon}_{ik}^2]}},$$

where $\widehat{\varepsilon}_{ik} = y_{ik} - \bar{y}_k$, $\bar{y}_k = \mathbb{E}_n[y_{ik}]$, and $(e_i)_{i=1}^n$ is a sequence of i.i.d. $N(0,1)$ random variables independent of the data. Theorem 19 implies that, when H_k are true for all k, $\mathrm{P}(\max_{1 \leqslant k \leqslant p} t_k > c_{1-\alpha}) \leqslant \alpha + o(1)$ (indeed, the inequality "$\leqslant$" can be replaced by the equality "$=$") uniformly in the underlying distribution provided that $\underline{\sigma}^2 \leqslant \mathrm{E}[\varepsilon_{ik}^2] \leqslant \bar{\sigma}^2$, $\log p = o(n^{1/7})$ and either (a) $\mathrm{E}[\max_{1 \leqslant k \leqslant p} \varepsilon_{ik}^4] \leqslant C_1$ or (b) $\mathrm{E}[\exp(|\varepsilon_{ik}|/C_1)] \leqslant 2$. Hence the multiplier bootstrap as described above leads to an asymptotically exact *testing procedure for the multiple hypothesis testing problem of which the* logarithm *of the number of hypotheses is nearly of order $n^{1/7}$ (subject to the prescribed assumptions). Note here that no assumption on the dependency structure between $y_{i1}, \dots, y_{ip}$ is made.*

The question on how large p can be was studied in [45] but from a conservative perspective. The motivation there is to know how fast p can grow to maintain the size of the simultaneous test when we calculate critical values (conservatively) ignoring the dependency among t_k and assuming that t_k were distributed as, say, $N(0,1)$. This framework is conservative in that correlation amongst statistics is dealt away with union bounds, namely by Bonferroni-Holm procedures. In contrast, our approach takes into account the correlation amongst statistics and hence is asymptotically exact, that is, asymptotically non-conservative. □

3.12 Appendix H. Application: Adaptive Specification Testing

In this section, we study the problem of adaptive specification testing. Let $(v_i, y_i)_{i=1}^n$ be a sample of independent random pairs where y_i is a scalar dependent random variable, and $v_i \in \mathbb{R}^d$ is a vector of non-stochastic covariates. The null hypothesis, H_0, is that there exists $\beta \in \mathbb{R}^d$ such that

$$\mathrm{E}[y_i] = v_i'\beta; \; i = 1, \dots, n. \tag{3.30}$$

The alternative hypothesis, H_a, is that there is no β satisfying (3.30). We allow for triangular array asymptotics so that everything in the model may depend on n. For brevity, however, we omit index n.

Let $\varepsilon_i = y_i - \mathrm{E}[y_i]$, $i = 1, \dots, n$. Then $\mathrm{E}[\varepsilon_i] = 0$, and under H_0, $y_i = v_i'\beta + \varepsilon_i$. To test H_0, consider a set of test functions $P_j(v_i)$, $j = 1, \dots, p$. Let $z_{ij} = P_j(v_i)$. We choose test functions so that $\mathbb{E}_n[z_{ij}v_i] = 0$ and $\mathbb{E}_n[z_{ij}^2] = 1$ for all $j = 1, \dots, p$. In our analysis, p may be higher or even much higher than n. Let $\widehat{\beta} = (\mathbb{E}_n[v_i v_i'])^{-1}(\mathbb{E}_n[v_i y_i])$ be an OLS estimator of β, and let $\widehat{\varepsilon}_i = y_i - z_i'\widehat{\beta}$; $i = 1, \dots, n$ be corresponding residuals. Our test statistic is

$$T := \max_{1 \leqslant j \leqslant p} \frac{|\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]}}.$$

The test rejects H_0 if T is significantly large.

Note that since $\mathbb{E}_n[z_{ij}v_i] = 0$, we have

$$\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i/\sqrt{n} = \sum_{i=1}^n z_{ij}(\varepsilon_i + v_i'(\beta - \widehat{\beta}))/\sqrt{n} = \sum_{i=1}^n z_{ij}\varepsilon_i/\sqrt{n}.$$

Therefore, under H_0,

$$T = \max_{1 \leqslant j \leqslant p} \frac{|\sum_{i=1}^n z_{ij}\varepsilon_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]}}.$$

This suggests that we can use the multiplier bootstrap to obtain a critical value for the test. More precisely, let $(e_i)_{i=1}^n$ be a sequence of independent $N(0,1)$ random variables that are independent of the data, and let

$$W := \max_{1 \leqslant j \leqslant p} \frac{|\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]}}.$$

The multiplier bootstrap critical value $c_W(1-\alpha)$ is the conditional $(1-\alpha)$-quantile of W given the data. To prove the validity of multiplier bootstrap, we will impose the following condition:

(S) There are some constants $c_1 > 0, C_1 > 0, \bar{\sigma}^2 > 0, \underline{\sigma}^2 > 0$, and a sequence $B_n \geqslant 1$

of constants such that for all $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$, $1 \leqslant k \leqslant d$: (i) $|z_{ij}| \leqslant B_n$; (ii) $\mathbb{E}_n[z_{ij}^2] = 1$; (iii) $\underline{\sigma}^2 \leqslant \mathrm{E}[\varepsilon_i^2] \leqslant \bar{\sigma}^2$; (iv) $|v_{ik}| \leqslant C_1$; (v) $d \leqslant C_1$; and (vi) the minimum eigenvalue of $\mathbb{E}_n[v_i v_i']$ is bounded from below by c_1.

Theorem 20 (Size Control of Adaptive Specification Test). *Let $c_2 > 0$ be some constant. Suppose that assumption S is satisfied. Moreover, suppose that either*

(a) $\mathrm{E}[\varepsilon_i^4] \leqslant C_1$ *for all* $1 \leqslant i \leqslant n$ *and* $B_n^4(\log(pn))^7/n \leqslant C_1 n^{-c_2}$*; or*

(b) $\mathrm{E}[\exp(|\varepsilon_i|/C_1)] \leqslant 2$ *for all* $1 \leqslant i \leqslant n$ *and* $B_n^2(\log(pn))^7/n \leqslant C_1 n^{-c_2}$.

Then there exist constants $c > 0$ and $C > 0$, depending only on $c_1, c_2, C_1, \underline{\sigma}^2$ and $\bar{\sigma}^2$, such that under H_0, $|\mathrm{P}(T \leqslant c_W(1-\alpha)) - (1-\alpha)| \leqslant Cn^{-c}$.

Comment 28. *The literature on specification testing is large. In particular, [63] and [54] developed adaptive tests that are suitable for inference in L_2-norm. In contrast, our test is most suitable for inference in* sup*-norm. An advantage of our procedure is that selecting a wide class of test functions leads to a test that can effectively adapt to a wide range of alternatives, including those that can not be well-approximated by Hölder-continuous functions.* □

3.13 Appendix I. Proofs for Section 3.11

3.13.1 Proof of Theorem 19

The multiplier bootstrap critical value $c_{1-\alpha,w}$ clearly satisfies $c_{1-\alpha,w} \leqslant c_{1-\alpha,w'}$ whenever $w \subset w'$, so inequality (3.27) is satisfied. Therefore, it suffices to prove (3.28). For the notational convenience, we will only consider the $w = \mathcal{W}$ case and suppress the uniformity in the underlying distribution. The general case follows from inspection of the proof.

Let us define

$$T := \max_{k,j} \frac{|\sum_{i=1}^n x_{ikj}\varepsilon_{ik}/\sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2 \widehat{\varepsilon}_{ik}^2]}}, \quad W := \max_{k,j} \frac{|\sum_{i=1}^n x_{ikj}\widehat{\varepsilon}_{ik}e_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2 \widehat{\varepsilon}_{ik}^2]}}.$$

We shall prove that $\mathrm{P}(T > c_W(1-\alpha)) = \alpha + o(1)$, where recall that $c_W(1-\alpha)$ is the conditional $(1-\alpha)$-quantile of W given (ε_{ik}). Here we will only consider case (a) of the theorem. The proof for case (b) is similar and hence omitted.

We make use of Corollary 6-(iv) to prove the desired claim. Define

$$T_0 := \max_{k,j} \frac{|\sum_{i=1}^n x_{ikj}\varepsilon_{ik}/\sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2 \sigma_i^2]}}, \; W_0 := \max_{k,j} \frac{|\sum_{i=1}^n x_{ikj}\varepsilon_{ik}e_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2 \sigma_{ik}^2]}}.$$

We first verify conditions (3.14) and (3.15) in Section 3.3. We will use the following facts directly deduced from assumption M:

$$\begin{aligned} \max_{i,k,j} |x_{ikj}| &\leqslant \max_{i,k} \|x_{ik}\| \leqslant_{(1)} c_1^{-1} \max_{i,k} \|v_{ik}\| \\ &\leqslant c_1^{-1}\sqrt{\bar{p}} \max_{i,k,j} |v_{ikj}| \leqslant_{(2)} c_1^{-1}\sqrt{\bar{p}} B_n, \qquad (3.31) \\ \max_{k,j} \mathbb{E}_n[x_{ikj}^2] &\leqslant \max_k \mathbb{E}_n[\|x_{ik}\|^2] \\ &\leqslant_{(3)} c_1^{-2} \max_k \mathbb{E}_n[\|v_{ik}\|^2] \leqslant_{(4)} c_1^{-2}\bar{p}, \qquad (3.32) \end{aligned}$$

where (1) and (3) follow from assumption M-(iv) and definition of x_{ik}, (2) is from M-(i) since v_{ik} is a subvector of z_i, and (4) is due to M-(ii). We shall first prove some lemmas. In these lemmas, we will assume all the conditions in Theorem 19 case (a) without mentioning so.

Lemma 48. $\sum_{i=1}^n x_{ikj}\varepsilon_{ik}/\sqrt{n} = O_{\mathrm{P}}(r_{n1})$ *uniformly over* $k = 1, \ldots, K$ *and* $j = 1, \ldots, p_k$ *where* $r_{n1} = \sqrt{\bar{p}\log p}$.

Proof. By Lemma 37 combined with inequalities (3.31) and (3.32), we have

$$\mathrm{E}[\max_{k,j} |\textstyle\sum_{i=1}^n x_{ikj}\varepsilon_{ik}/\sqrt{n}|] = O(\sqrt{\bar{p}}B_n(\log p)/n^{1/4} + \sqrt{\bar{p}\log p}) = O(\sqrt{\bar{p}\log p}),$$

where the second step follows because $B_n\sqrt{\log p}/n^{1/4} = o(1)$. The claim follows from Markov's inequality. □

Lemma 49. $\mathbb{E}_n[x_{ikj}^2(\hat{\varepsilon}_{ik}^2 - \sigma_{ik}^2)] = O_{\mathrm{P}}(r_{n2})$ *uniformly over* $k = 1, \ldots, K$ *and* $j = 1, \ldots, p_k$ *where* $r_{n2} = \bar{p}B_n^2(\log p)/\sqrt{n}$.

Proof. We have

$$\begin{aligned}\mathbb{E}_n[x_{ikj}^2(\widehat{\varepsilon}_{ik}^2 - \sigma_{ik}^2)] &= \mathbb{E}_n[x_{ikj}^2(\varepsilon_{ik}^2 - \sigma_{ik}^2)] + \mathbb{E}_n[x_{ikj}^2(v_{ik}'(\widehat{\beta}_k - \beta_k))^2] \\ &\quad - 2\mathbb{E}_n[x_{ikj}^2\varepsilon_{ik}v_{ik}'(\widehat{\beta}_k - \beta_k)] \\ &=: I_{jk} + II_{jk} + III_{jk}.\end{aligned}$$

We will show in steps 1-3 below that $I_{jk} = O_{\mathrm{P}}(\bar{p}B_n^2(\log p)/\sqrt{n})$, $II_{jk} = O_{\mathrm{P}}(\bar{p}^2B_n^2(\log p)/n)$, and $III_{jk} = O_{\mathrm{P}}(\bar{p}^2B_n^2(\log p)/n)$ uniformly over $k = 1, \dots, K$ and $j = 1, \dots, p_k$. The claim of the lemma follows since $\bar{p}/\sqrt{n} \to 0$.

Step 1. We prove that $I_{jk} = \mathbb{E}_n[x_{ikj}^2(\varepsilon_{ik}^2 - \sigma_{ik}^2)] = O_{\mathrm{P}}(\bar{p}B_n^2(\log p)/\sqrt{n})$ uniformly over $k = 1, \dots, K$ and $j = 1, \dots, p_k$.

By Lemma 37 combined with inequalities (3.31) and (3.32), we have

$$\begin{aligned}\mathrm{E}[\max_{k,j} |\mathbb{E}_n[x_{ikj}^2(\varepsilon_{ik}^2 - \sigma_{ik}^2)]|] &= O(\bar{p}B_n^2(\log p)/\sqrt{n} + \bar{p}B_n\sqrt{(\log p)/n}) \\ &= O(\bar{p}B_n^2(\log p)/\sqrt{n}),\end{aligned}$$

where the second step follows because $B_n \geqslant 1$. The claim of this step follows from Markov's inequality.

Step 2. We prove that

$$II_{jk} = \mathbb{E}_n[x_{ikj}^2(v_{ik}'(\widehat{\beta}_k - \beta_k))^2] = O_{\mathrm{P}}(\bar{p}^2B_n^2(\log p)/n)$$

uniformly over $k = 1, \dots, K$ and $j = 1, \dots, p_k$. We have

$$\begin{aligned}\max_{k,j} \mathbb{E}_n[x_{ikj}^2(v_{ik}'(\widehat{\beta}_k - \beta_k))^2] &\leqslant_{(1)} c_1^{-2}\bar{p}B_n^2 \max_k \mathbb{E}_n[(v_{ik}'(\widehat{\beta}_k - \beta_k))^2] \\ &= c_1^{-2}\bar{p}B_n^2 \max_k \mathbb{E}_n[\varepsilon_{ik}v_{ik}']\mathbb{E}_n[v_{ik}v_{ik}']^{-1}\mathbb{E}_n[v_{ik}\varepsilon_{ik}] \\ &\leqslant_{(2)} c_1^{-3}\bar{p}B_n^2 \max_k \|\mathbb{E}_n[v_{ik}\varepsilon_{ik}]\|^2 \\ &\leqslant c_1^{-3}\bar{p}^2B_n^2 \max_{k,j} |\mathbb{E}_n[v_{ikj}\varepsilon_{ik}]|^2 \\ &=_{(3)} O_{\mathrm{P}}(\bar{p}^2B_n^2(\log p)/n),\end{aligned}$$

where (1) follows from inequality (3.31), (2) from assumption M-(iv), and (3) from application of Lemma 37. The claim of this step follows.

Step 3. We prove that

$$III_{jk} = \mathbb{E}_n[x_{ikj}^2 \varepsilon_{ik}(v_{ik}'(\widehat{\beta}_k - \beta_k))] = O_{\mathrm{P}}(\bar{p}^2 B_n^2 (\log p)/n)$$

uniformly over $k = 1, \ldots, K$ and $j = 1, \ldots, p_k$. We have

$$\begin{aligned}\max_{k,j} |\mathbb{E}_n[x_{ikj}^2 \varepsilon_{ik}(v_{ik}'(\widehat{\beta}_k - \beta_k))]| &\leqslant \max_{k,j} \|\mathbb{E}_n[x_{ikj}^2 \varepsilon_{ik} v_{ik}']\| \|\widehat{\beta}_k - \beta_k\| \\ &\leqslant \max_{k,j,l} \sqrt{\bar{p}} |\mathbb{E}_n[x_{ikj}^2 \varepsilon_{ik} v_{ikl}]| \|\widehat{\beta}_k - \beta_k\|.\end{aligned}$$

Then

$$\begin{aligned}&\max_{k} \|\widehat{\beta}_k - \beta_k\| = \max_{k} \|\mathbb{E}_n[v_{ik} v_{ik}']^{-1} \mathbb{E}_n[v_{ik}\varepsilon_{ik}]\| \leqslant_{(1)} c_1^{-1} \max_{k} \|\mathbb{E}_n[v_{ik}\varepsilon_{ik}]\| \\ &\leqslant c_1^{-1} \sqrt{\bar{p}} \max_{k,j} |\mathbb{E}_n[v_{ikj}\varepsilon_{ik}]| =_{(2)} O_{\mathrm{P}}(\sqrt{\bar{p}(\log p)/n})\end{aligned}$$

where (1) follows from assumption M-(iv) and (2) is as in step 2. In addition, by Lemma 37 combined with inequalities (3.31) and (3.32), we have

$$\begin{aligned}\mathrm{E}[\max_{k,j,l} |\mathbb{E}_n[x_{ikj}^2 \varepsilon_{ik} v_{ikl}]|] &= O(\bar{p} B_n^3 (\log p)/n^{3/4} + \bar{p} B_n^2 \sqrt{(\log p)/n}) \\ &= O(\bar{p} B_n^2 \sqrt{(\log p)/n}),\end{aligned}$$

where the last step is because $B_n \sqrt{\log p}/n^{1/4} = o(1)$. Combining these bounds yields the claim of this step. □

In Lemmas 50 and 51, $\mathrm{E}_e[\cdot]$ denotes the expectation with respect to $(e_i)_{i=1}^n$ conditional on (ε_{ik}).

Lemma 50. *$\sum_{i=1}^n x_{ikj}\widehat{\varepsilon}_{ik} e_i/\sqrt{n} = O_{\mathrm{P}}(r_{n1})$ uniformly over $k = 1, \ldots, K$ and $j = 1, \ldots, p_k$. Recall that $r_{n1} = \sqrt{\bar{p}\log p}$.*

Proof. We have

$$
\begin{aligned}
&\mathrm{E}_e[\max_{k,j} |\textstyle\sum_{i=1}^n x_{ikj}\widehat{\varepsilon}_{ik}e_i/\sqrt{n}|] \lesssim_{(1)} \sqrt{\log p}\max_{k,j}(\mathbb{E}_n[x_{ikj}^2\widehat{\varepsilon}_{ik}^2])^{1/2}\\
&=_{(2)} \sqrt{\log p}\max_{k,j}(\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2]+O_{\mathrm{P}}(r_{n2}))^{1/2}\\
&\leqslant_{(3)} \sqrt{\log p}\max_{k,j}(\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2])^{1/2}+O_{\mathrm{P}}(r_{n2}\sqrt{\log p})\\
&\leqslant_{(4)} \sigma\sqrt{\log p}\max_{k,j}(\mathbb{E}_n[x_{ikj}^2])^{1/2}+O_{\mathrm{P}}(r_{n2}\sqrt{\log p})\\
&=_{(5)} O_{\mathrm{P}}(\sqrt{\bar{p}\log p}),
\end{aligned}
$$

where (1) follows from Pisier's inequality, (2) from lemma 49, (3) follows from application of Taylor's theorem together with the fact that $r_{n2} = o(1)$ and $\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2]$ is bounded away from zero (which is guaranteed by assumptions M-(iii) and M-(v)) (4) follows from assumption M-(iii), and (5) is due to equation (3.32) and $r_{n2} = o(1)$. The claim of the lemma follows. □

Lemma 51. $\sum_{i=1}^n x_{ikj}(\widehat{\varepsilon}_{ik}-\varepsilon_{ik})e_i/\sqrt{n} = O_{\mathrm{P}}(r_{n3})$ *uniformly over* $k = 1, \dots, K$ *and* $j = 1, \dots, p_k$ *where* $r_{n3} = \bar{p}B_n(\log p)/\sqrt{n}$.

Proof. We have

$$
\begin{aligned}
\mathrm{E}_e[|\textstyle\sum_{i=1}^n x_{ikj}(\widehat{\varepsilon}_{ik}-\varepsilon_{ik})e_i/\sqrt{n}|] &\lesssim_{(1)} \sqrt{\log p}\max_{k,j}(\mathbb{E}_n[x_{ikj}^2(\widehat{\varepsilon}_{ik}-\varepsilon_{ik})^2])^{1/2}\\
&=_{(2)} \sqrt{\log p}\max_{k,j}(\mathbb{E}_n[x_{ikj}^2(v_{ik}'(\widehat{\beta}_k-\beta_k))^2])^{1/2}\\
&=_{(3)} O_{\mathrm{P}}(\bar{p}B_n(\log p)/\sqrt{n})
\end{aligned}
$$

where (1) follows from Pisier's inequality, (2) is by definition of $\widehat{\varepsilon}_{ik}$, and (3) is by step 2 in the proof of lemma 49. The claim follows. □

Going back to the proof of Theorem 19, by Lemmas 48 and 49 and the fact that

$\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2]$ is bounded away from zero, we have

$$T = \max_{k,j} \frac{|\sum_{i=1}^n x_{ikj}\varepsilon_{ik}/\sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2] + O_{\mathrm{P}}(r_{n2})}}$$
$$= T_0 + O_{\mathrm{P}}(r_{n1}r_{n2}) = T_0 + o_{\mathrm{P}}(1/\sqrt{\log p}),$$

where the last step uses the fact that $\bar{p}^3 B_n^4 (\log p)^4/n = o(1)$. Similarly, by Lemmas 49-51, we have

$$W = \max_{k,j} \frac{|\sum_{i=1}^n x_{ikj}\widehat{\varepsilon}_{ik}e_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2]}} + O_{\mathrm{P}}(r_{n1}r_{n2})$$
$$= W_0 + O_{\mathrm{P}}(r_{n1}r_{n2} + r_{n3}) = W_0 + o_{\mathrm{P}}(1/\sqrt{\log p}),$$

where the last step uses the fact that $\bar{p}B_n(\log p)^{3/2}/\sqrt{n} = o(1)$. Hence it is verified that conditions (3.14) and (3.15) in Section 3.3 are satisfied with some sequences $\zeta_1 = \zeta_{1n} \to 0$ and $\zeta_2 = \zeta_{2n} \to 0$ such that $\zeta_1\sqrt{\log p} + \zeta_2 = o(1)$. Therefore, the desired claim follows from Corollary 6-(iv). □

3.14 Appendix J. Proofs for Section 3.12

3.14.1 Proof of Theorem 20

We only consider case (a). The proof for case (b) is similar and hence omitted. In this proof, let c, c', C, C' denote generic positive constants depending only on $c_1, c_2, C_1, \underline{\sigma}^2, \bar{\sigma}^2$ and their values may change from place to place. Let

$$T_0 := \max_{1\leqslant j\leqslant p} \frac{|\sum_{i=1}^n z_{ij}\varepsilon_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2\sigma_i^2]}} \text{ and } W_0 := \max_{1\leqslant j\leqslant p} \frac{|\sum_{i=1}^n z_{ij}\varepsilon_i e_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2\sigma_i^2]}}.$$

We make use of Corollary 6-(iv). To this end, we shall verify conditions (3.14) and (3.15) in Section 3.3, which will be separately done in Steps 1 and 2, respectively.

Step 1. We show that $\mathrm{P}(|T - T_0| > \zeta_1) < \zeta_2$ for some ζ_1 and ζ_2 satisfying

$\zeta_1\sqrt{\log p} + \zeta_2 \leqslant Cn^{-c}$.

By Corollary 5-(v), we have

$$\begin{aligned} &\mathrm{P}\left(\max_{1\leqslant j\leqslant p} |\textstyle\sum_{i=1}^n z_{ij}\varepsilon_i/\sqrt{n}| > t\right) \\ &\quad \leqslant \mathrm{P}\left(\max_{1\leqslant j\leqslant p} |\textstyle\sum_{i=1}^n z_{ij}\sigma_i e_i/\sqrt{n}| > t\right) + Cn^{-c}, \end{aligned}$$

uniformly in $t \in \mathbb{R}$. By the Gaussian Concentration Inequality, for every $t > 0$, we have

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p} |\textstyle\sum_{i=1}^n z_{ij}\sigma_i e_i/\sqrt{n}| > \mathrm{E}[\max_{1\leqslant j\leqslant p} |\textstyle\sum_{i=1}^n z_{ij}\sigma_i e_i/\sqrt{n}|] + Ct\right) \leqslant e^{-t^2}.$$

Since $\mathrm{E}[\max_{1\leqslant j\leqslant p} |\sum_{i=1}^n z_{ij}\sigma_i e_i/\sqrt{n}|] \leqslant C\sqrt{\log p}$, we conclude that

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p} |\textstyle\sum_{i=1}^n z_{ij}\varepsilon_i/\sqrt{n}| > C\sqrt{\log(pn)}\right) \leqslant C'n^{-c}. \tag{3.33}$$

Moreover,

$$\begin{aligned} \mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2 - \sigma_i^2)] &= \mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2] + \mathbb{E}_n[z_{ij}^2(\varepsilon_i^2 - \sigma_i^2)] + 2\mathbb{E}_n[z_{ij}^2\varepsilon_i(\widehat{\varepsilon}_i - \varepsilon_i)] \\ &=: I_j + II_j + III_j. \end{aligned}$$

Consider I_j. We have

$$I_j \leqslant_{(1)} \max_{1\leqslant i\leqslant n}(\widehat{\varepsilon}_i - \varepsilon_i)^2 \leqslant_{(2)} C\|\widehat{\beta} - \beta\|^2 \leqslant_{(3)} C'\|\mathbb{E}_n[v_i\varepsilon_i]\|^2,$$

where (1) follows from assumption S-(ii), (2) from S-(iv) and S-(v), and (3) from S-(vi). Since $\mathrm{E}[\|\mathbb{E}_n[v_i\varepsilon_i]\|^2] \leqslant C/n$, by Markov's inequality, for every $t > 0$,

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p} \mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2] > t\right) \leqslant C/(nt). \tag{3.34}$$

Consider II_j. By Lemma 37 and Markov's inequality, we have

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2(\varepsilon_i^2-\sigma_i^2)]|>t\right)\leqslant CB_n^2(\log p)/(\sqrt{n}t). \tag{3.35}$$

Consider III_j. We have $|III_j|\leqslant 2|\mathbb{E}_n[z_{ij}^2 v_i'(\beta-\widehat{\beta})\varepsilon_i]|\leqslant 2\|\mathbb{E}_n[z_{ij}^2\varepsilon_i v_i]\|\|\widehat{\beta}-\beta\|$. Hence

$$\begin{aligned}
&\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2\varepsilon_i(\widehat{\varepsilon}_i-\varepsilon_i)]|>t\right)\\
&\leqslant \mathrm{P}\left(\max_{1\leqslant j\leqslant p}\|\mathbb{E}_n[z_{ij}^2\varepsilon_i v_i]\|>t\right)+\mathrm{P}(\|\widehat{\beta}-\beta\|>1)\\
&\leqslant C[B_n^2(\log p)/(\sqrt{n}t)+1/n]. \qquad (3.36)
\end{aligned}$$

By (3.34)-(3.36), we have

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2-\sigma_i^2)]|>t\right)\leqslant C[B_n^2(\log p)/(\sqrt{n}t)+1/(nt)+1/n]. \tag{3.37}$$

In particular,

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2-\sigma_i^2)]|>\underline{\sigma}^2/2\right)\leqslant Cn^{-c}.$$

Since $\mathbb{E}_n[z_{ij}^2\sigma_i^2]\geqslant\underline{\sigma}^2>0$ (which is guaranteed by S-(iii) and S-(ii)), on the event $\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2-\sigma_i^2)]|\leqslant\underline{\sigma}^2/2$, we have

$$\min_{1\leqslant j\leqslant p}\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]\geqslant\min_{1\leqslant j\leqslant p}\mathbb{E}_n[z_{ij}^2\sigma_i^2]-\underline{\sigma}^2/2\geqslant\underline{\sigma}^2/2,$$

and hence

$$\begin{aligned}
|T-T_0|&=\max_{1\leqslant j\leqslant p}\left|\frac{\sqrt{\mathbb{E}_n[z_{ij}^2\sigma_i^2]}-\sqrt{\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]}}{\sqrt{\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]}}\right|\times T_0\\
&\leqslant C\max_{1\leqslant j\leqslant p}\left|\sqrt{\mathbb{E}_n[z_{ij}^2\sigma_i^2]}-\sqrt{\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]}\right|\times T_0\\
&\leqslant C\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2\sigma_i^2]-\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]|\times T_0,
\end{aligned}$$

where the last step uses the simple fact that

$$|\sqrt{a}-\sqrt{b}|=\frac{|a-b|}{\sqrt{a}+\sqrt{b}}\leqslant\frac{|a-b|}{\sqrt{a}}.$$

By (3.33) and (3.37), for every $t>0$,

$$\mathrm{P}\left(|T-T_0|>Ct\sqrt{\log(pn)}\right)\leqslant C'[n^{-c}+B_n^2(\log p)/(\sqrt{n}t)+1/(nt)].$$

By choosing $t=(\log(pn))^{-1}n^{-c'}$ with sufficiently small $c'>0$, we obtain the claim of this step.

Step 2. We show that $\mathrm{P}(\mathrm{P}_e(|W-W_0|>\zeta_1)>\zeta_2)<\zeta_2$ for some ζ_1 and ζ_2 satisfying $\zeta_1\sqrt{\log p}+\zeta_2\leqslant Cn^{-c}$.

For $0<t\leqslant\underline{\sigma}^2/2$, consider the event

$$\mathcal{E}=\left\{(\varepsilon_i)_{i=1}^n:\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2-\sigma_i^2)]|\leqslant t,\max_{1\leqslant i\leqslant p}(\widehat{\varepsilon}_i-\varepsilon_i)^2\leqslant t^2\right\}.$$

By calculations in Step 1, $\mathrm{P}(\mathcal{E})\geqslant 1-C[B_n^2(\log p)/(\sqrt{n}t)+1/(nt^2)+1/n]$. We shall show that, on this event,

$$\mathrm{P}_e\left(\max_{1\leqslant j\leqslant p}|\textstyle\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}|>C\sqrt{\log(pn)}\right)\leqslant n^{-1}, \tag{3.38}$$

$$\mathrm{P}_e\left(\max_{1\leqslant j\leqslant p}|\textstyle\sum_{i=1}^n z_{ij}(\widehat{\varepsilon}_i-\varepsilon_i)e_i/\sqrt{n}|>Ct\sqrt{\log(pn)}\right)\leqslant n^{-1}. \tag{3.39}$$

For (3.38), by the Gaussian concentration inequality, for every $s>0$,

$$\mathrm{P}_e\left(\max_{1\leqslant j\leqslant p}|\textstyle\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}|>\mathrm{E}_e[\max_{1\leqslant j\leqslant p}|\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}|]+Cs\right)\leqslant e^{-s^2}.$$

where we have used the fact $\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]=\mathbb{E}_n[z_{ij}^2\sigma_i^2]+\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2-\sigma_i^2)]\leqslant\bar{\sigma}^2+t\leqslant\bar{\sigma}^2+\underline{\sigma}^2/2$ on the event $\mathcal{E}$. Here $\mathrm{E}_e[\cdot]$ means the expectation with respect to $(e_i)_{i=1}^n$ conditional on $(\varepsilon_i)_{i=1}^n$. Moreover, on the event $\mathcal{E}$,

$$\mathrm{E}_e[\max_{1\leqslant j\leqslant p}|\textstyle\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}|]\leqslant C\sqrt{\log p}.$$

Hence by choosing $s = \sqrt{\log n}$, we obtain (3.38). Inequality (3.39) follows similarly, by noting that $(\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2])^{1/2} \leqslant \max_{1 \leqslant i \leqslant n} |\widehat{\varepsilon}_i - \varepsilon_i| \leqslant t$ on the event $\mathcal{E}$.

Define

$$W_1 := \max_{1 \leqslant j \leqslant p} \frac{|\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2 \sigma_i^2]}}.$$

Note that $\mathbb{E}_n[z_{ij}^2\sigma_i^2] \geqslant \underline{\sigma}^2$. Since on the event $\mathcal{E}$, $\max_{1 \leqslant j \leqslant p} |\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2 - \sigma_i^2)]| \leqslant t \leqslant \underline{\sigma}^2/2$, in view of Step 1, on this event, we have

$$\begin{aligned}
|W - W_0| &\leqslant |W - W_1| + |W_1 - W_0| \\
&\leqslant CtW_1 + |W_1 - W_0| \\
&\leqslant Ct \max_{1 \leqslant j \leqslant p} |\textstyle\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}| + C \max_{1 \leqslant j \leqslant p} |\textstyle\sum_{i=1}^n z_{ij}(\widehat{\varepsilon}_i - \varepsilon_i)e_i/\sqrt{n}|.
\end{aligned}$$

Therefore, by (3.38) and (3.39), on the event $\mathcal{E}$, we have

$$\mathrm{P}_e\left(|W - W_0| > Ct\sqrt{\log(pn)}\right) \leqslant 2n^{-1}.$$

By choosing $t = (\log(pn))^{-1}n^{-c}$ with sufficiently small $c > 0$, we obtain the claim of this step.

Step 3. Steps 1 and 2 verified conditions (3.14) and (3.15) in Section 3.3. Theorem 20 case (a) follows from Corollary 6-(iv). □

Bibliography

[1] P. Alquier and M. Hebiri. Generalization of ℓ_1 constraints for high dimensional regression problems. *Statist. Probab. Lett.*, 81:1760–1765, 2011.

[2] E. Anderson and D. Schmittlein. Integration of the sales force: An empirical examination. *RAND Journal of Economics*, 15:385–395, 1984.

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

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

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

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

[7] M. Angeletos and I. Werning. Information aggregation, multiplicity, and volatility. *American Economic Review*, 96(5), 2006.

[8] S. Arlot, G. Blanchard, and E. Roquain. Some non-asymptotic results on resampling in high dimension i: confidence regions. *Ann. Statist.*, 38:83–99, 2010.

[9] S. Arlot, G. Blanchard, and E. Roquain. Some non-asymptotic results on resampling in high dimension ii: multiple tests. *Ann. Statist.*, 38:83–99, 2010.

[10] T. Armstrong. Asymptotically exact inference in conditional moment inequalities models. *unpublished manuscript*, 2011.

[11] T. Armstrong. Weighted ks statistics for inference on conditional moment inequalities. *unpublished manuscript*, 2011.

[12] T. Armstrong and H. P. Chan. Multiscale adaptive inference on conditional moment inequalities. *arXiv:1212.5729*, pages 1–47, 2012.

[13] S. Athey. Monotone comparative statics under uncertainty. *The Quarterly Journal of Economics*, 117:187–223, 2002.

[14] K. Ball. The reverse isoperimetric problem for gaussian measure. *Discrete Comput. Geom.*, 10:411–420, 1993.

[15] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, forthcoming.

[16] A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98:791–806, 2011.

[17] V. Bentkus. On the dependence of the berry-esseen bound on dimension. *J. Statist. Plann. Infer.*, 113:385–402, 2003.

[18] Yannick Beraud, Sylvie Huet, and Beatrice Laurent. 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, 2005.

[19] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37:1705–1732, 2009.

[20] S. Boucheron, O. Bousquet, and G. Lugosi. Concentration inequalities. *Advanced Lectures in Machine Learning*, pages 208–240, 2004.

[21] A. W. Bowman, M. C. Jones, and Irene Gijbels. Testing monotonicity of regression. *Journal of Computational and Graphical Statistics*, 7:489–500, 1998.

[22] J. Bretagnolle and P. Massart. Hungarian construction from the non asymptotic viewpoint. *Ann. Probab.*, 17:239–256, 1989.

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

[24] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods,Theory and Applications.* Springer, 2011.

[25] T. Cai and L. Wang. Adaptive variance function estimation in heterscedastic nonparametric regression. *The Annals of Statistics*, 36:2025–2054, 2008.

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

[27] E.J. Candès and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35:2313–2351, 2007.

[28] S. Chatterjee. An error bound in the sudakov-fernique inequality. *arXiv:math/0510424.*

[29] S. Chatterjee and E. Meckes. Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.*, 4:257–283, 2008.

[30] L. Chen and X. Fang. Multivariate normal approximation by stein's method: the concentration inequality approach. *arXiv:1111.4073.*

[31] L. Chen, L. Goldstein, and Q.-M. Shao. *Normal Approximation by Stein's Method.* Springer, 2011.

[32] V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximation of suprema of empirical processes. *arXiv:1212.6906.*

[33] V. Chernozhukov, D. Chetverikov, and K. Kato. Central limit theorems and multiplier bootstrap when p is much larger than n. *aRxiv:1212.6906v3*, pages 1–49, 2012.

[34] V. Chernozhukov, D. Chetverikov, and K. Kato. Comparison and anti-concentration bounds for maxima of gaussian random vectors. *arXiv:1301.4807v3*, pages 1–21, 2013.

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

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

[37] D. Chetverikov. Adaptive test of conditional moment inequalities. *arXiv:1201.0167v2*, 2011.

[38] D. Chetverikov. Testing regression monotonicity in econometric models. *arXiv:1212.6885*, 2012.

[39] V. de la Peña, T. Lai, and Q.-M. Shao. *Self-Normalized Processes: Limit Theory and Statistical Applications.* Springer, 2009.

[40] M. Delgado and J. Escanciano. Testing conditional monotonicity in the absence of smoothness. *working paper*, pages 1–18, 2010.

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

[42] L. Dumbgen and V. G. Spokoiny. Multiscale testing of qualitative hypotheses. *The Annals of Statistics*, 29:124–152, 2001.

[43] Cecile Durot. A kolmogorov-type test for monotonicity of regression. *Statistics and Probability Letters*, 63:425–433, 2003.

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

[45] J. Fan, P. Hall, and Q. Yao. To how many simultaneous hypothesis tests can normal, student's t or bootstrap calibration be applied. *J. Amer. Stat. Assoc.*, 102:1282–1288, 2007.

[46] J. Fan and Q. Yao. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85:645–660, 1998.

[47] K. Frick, P. Marnitz, and A. Munk. Shape-constrained regularization by statistical multiresolution for inverse problems: asymptotic analysis. *Inverse Problems*, 28, 2012.

[48] E. Gautier and A. Tsybakov. High-dimensional istrumental variables regression and confidence sets. *arXiv: 1105.2454*.

[49] Subhashis Ghosal, Arusharka Sen, and Aad van der Vaart. Testing monotonicity of regression. *The Annals of Statistics*, 28:1054–1082, 2000.

[50] Irene Gijbels, Peter Hall, M. C. Jones, and Inge Koch. Tests for monotonicity of a regression mean with guaranteed level. *Biometrika*, 87:663–673, 2000.

[51] E. Gine and R. Nickl. Confidence bands in density estimation. *The Annals of Statistics*, 38:1120–1170, 2010.

[52] L. Goldstein and Y. Rinott. Multivariate normal approximations by stein's method and size bias couplings. *J. Appl. Probab.*, 33:1–17, 1996.

[53] F. Götze. On the rate of convergence in the multivariate clt. *Ann. Probab.*, 19:724–739, 1991.

[54] E. Guerre and P. Lavergne. Data-driven rate-optimal specification testing in regression models. *Ann. Statist.*, 33:840–870, 2005.

[55] P. Haile and E. Tamer. Inference with an incomplete model of english auctions. *Journal of Political Economy*, 111:1–51, 2003.

[56] P. Hall. On convergence rates of suprema. *Probability Theory and Related Fields*, 89:447–455, 1991.

[57] Peter Hall and Nancy Heckman. Testing for monotonicity of a regression mean by calibrating for linear functions. *The Annals of Statistics*, 28:20–39, 2000.

[58] W. Hardle and E. Mammen. Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 21:1926–1947, 1993.

[59] W. Hardle and A. Tsybakov. Local polinomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics*, 81:233–242, 1997.

[60] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

[61] B. Holmstrom and P. Milgrom. The firm as an incentive system. *The American Economic Review*, 84:972–991, 1994.

[62] J. L. Horowitz. *Semiparametric and Nonparametric Methods in Econometrics.* Springer Series in Statistics, 2009.

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

[64] A. Juditsky and A. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. *Math. Program. Ser. B*, 127:57–88, 2011.

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

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

[67] V. Koltchinskii. Komlós-major-tusnády approximation for the general empirical process and haar expansions of classes of functions. *J. Theoret. Probab.*, 7:73–118, 1994.

[68] V. Koltchinskii. The dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15:799–828, 2009.

[69] J. Komlós, P. Major, and G. Tusnády. An approximation for partial sums of independent rv's and the sample df i. *Z. Warhsch. Verw. Gabiete*, 32:111–131, 1975.

[70] R. Kress. *Linear Integral Equations.* Springer, 1999.

[71] M. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and Related Properties of Random Sequences and Processes.* Springer, 1983.

[72] S. Lee, O. Linton, and Y Whang. Testing for stochastic monotonicity. *Econometrica*, 77(2):585–602, 2009.

[73] S. Lee, K. Song, and Y. J. Whang. Nonparametric tests of monotonicity: an l_p approach. *working paper*, 2011.

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

[75] E.L. Lehmann and J.P. Romano. *Testing Statistical Hypotheses.* Springer, 2005.

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

[77] R. Liu. Bootstrap procedures under niid models. *The Annals of Statistics*, 16(4):1696–1708, 1988.

[78] Enno Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, 21(1):255–285, 1993.

[79] C. F. Manski and V. Pepper, J. Monotone instrumental variables: With an application to returns to schooling. *Econometrica*, 68:997–1010, 2000.

[80] C. F. Manski and E. Tamer. Inference on regressions with interval data on a regressor or outcome. *Econometrica*, 70:519–546, 2002.

[81] R. Merton. On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance*, 29:449–470, 1974.

[82] P. Milgrom and C. Shannon. Monotone comparative statics. *Econometrica*, 62:157–180, 1994.

[83] P. Milgrom and R. Weber. A theory of auctions and competitive bidding. *Econometrica*, 50:1089–1122, 1982.

[84] S. Morris and S. Shin, H. Unique equilibrium in a model of self-fulfilling currency attacks. *The American Economic Review*, 88(3):587–597, 1998.

[85] S. Morris and S. Shin, H. Global games: Theory and application. *Cowles Foundation Discussion Paper, No 1275R*, pages 1–70, 2001.

[86] S. Morris and S. Shin, H. Coordination risk and the price of debt. *European Economic Review*, 48:133–153, 2004.

[87] H-G Muller. Smooth optimum kernel estimators near endpoints. *Biometrika*, 78(3):521–30, 1991.

[88] H. G. Muller and U. Stadtmuller. Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics*, 15:610–625, 1987.

[89] S. Nagaev. An estimate of the remainder term in the multidimensional central limit theorem. *Proc. Third Japan-USSR Symp. Probab. Theory*, pages 419–438, 1976.

[90] W. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79:147–168, 1997.

[91] A. Pakes. Alternative models for moment inequalities. *Econometrica*, 78:1783–1822, 2010.

[92] D. Panchenko. *The Sherrington-Kirkpatrick Model*. Springer, 2013.

[93] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[94] D. Pollard. *A User's Guide to Measure Theoretic Probability.* Cambridge University Press, 2002.

[95] L. Poppo and T. Zenger. 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, 1998.

[96] S. Portnoy. On the central limit theorem in $\mathbb{R}^p$ when $p \to \infty$. *Probab. Theory Related Fields*, 73:571–583, 1986.

[97] J. Rice. Bandwidth choice for nonparametric kernel regression. *The Annals of Statistics*, 12:1215–1230, 1984.

[98] E. Rio. Local invariance principles and their application to density estimation. *Probability Theory and Related Fields*, 98:21–45, 1994.

[99] P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56:931–954, 1988.

[100] J. Romano and M. Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005.

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

[102] P. Romano, J. and A. M. Shaikh. Inference for the identified sets in partially identified econometric models. *Econometrica*, 78:169–211, 2010.

[103] P. Romano, J. and M. Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100:94–108, 2005.

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

[105] W. Schlee. Nonparametric tests of the monotony and convexity of regression. In *Nonparametric Statistical Inference.* Amsterdam: North-Holland, 1982.

[106] D. Slepian. The one-sided barrier problem for gaussian noise. *Bell Syst. Tech. J.*, 41:463–501, 1962.

[107] C. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9:1135–1151, 1981.

[108] M. Talagrand. *Spin Glasses: A Challenge for Mathematicians.* Springer, 2003.

[109] Jean Tirole. *The Theory of Industrial Organization.* Cambridge, MA: MIT Press, 1988.

[110] A. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2009.

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

[112] Jianqiang Wang and Mary Meyer. Testing the monotonicity or convexity of a function using regression splines. *The Canadian Journal of Statistics*, 39:89–107, 2011.

[113] C.F.J. Wu. Jacknife, bootstrap, and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, 1986.

[114] F. Ye and C. Zhang. Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.*, 11:3519–3540, 2010.