

# Multiscale Adaptive Inference on Conditional Moment Inequalities

Timothy B. Armstrong

Yale University \*

Hock Peng Chan

National University of Singapore †

December 22, 2012

## Abstract

This paper considers inference for conditional moment inequality models using a multiscale statistic. We derive the asymptotic distribution of this test statistic and use the result to propose feasible critical values that have a simple analytic formula. We also propose critical values based on a modified bootstrap procedure and prove their asymptotic validity. The asymptotic distribution is extreme value, and the proof uses new techniques to overcome several technical obstacles. We provide power results that show that our test detects local alternatives that approach the identified set at the best possible rate under a set of conditions that hold generically in the set identified case in a broad class of models, and that our test is adaptive to the smoothness properties of the data generating process. Our results also have implications for the use of moment selection procedures in this setting. We provide a monte carlo study and an empirical illustration to inference in a regression model with endogenously censored and missing data.

---

We thank Tze Leung Lai, Joe Romano, Han Hong, Don Andrews, Xiaohong Chen and seminar participants at Yale, UT Austin, Northwestern and the 2012 Greater New York Metropolitan Econometrics Colloquium for helpful comments and Bonny Wang for research assistance. Armstrong acknowledges support from a fellowship from the endowment in memory of B.F. Haley and E.S. Shaw through SIEPR.

\*email: timothy.armstrong@yale.edu

†email: stachp@nus.edu.sg

# 1 Introduction

This paper shows how to achieve optimal power adaptively in conditional moment inequality models using a multiscale test statistic. Formally, the model is defined by a vector of inequality restrictions of the form  $E(m(W_i, \theta)|X_i) \geq 0$  almost surely, where  $m$  is a known parametric function and inequality is taken elementwise. The object of interest is the identified set  $\Theta_0$  of parameter values that satisfy this set of restrictions, and the goal is to form a test that has good power properties at alternative values of  $\theta$  near the boundary of the identified set. By testing the null  $\theta \in \Theta_0$  for each  $\theta$ , and inverting these tests, one obtains a confidence region that, for each point in the identified set, contains this point with a pre-specified probability (see Imbens and Manski, 2004, for a discussion of this and other notions of inference in this setting). This class of models includes numerous models used in empirical economics, including selection models, regression models with endogenously missing or censored data, and certain models of firm and consumer behavior (see below for references from the literature).

We derive the asymptotic distribution of our test statistic and show how it can be used to obtain feasible critical values. We provide two methods for computing critical values for our test statistic and prove the validity of both using our asymptotic distribution result. The first is based on the asymptotic distribution itself and has the advantage of having a simple analytic formula that can be computed without using simulation. This is particularly useful in applied settings where computational issues can severely limit the applicability of tests that require resampling or simulation to compute critical values. The second method for computing critical values uses a modified bootstrap procedure. While we focus on least favorable critical values, both methods can be used with first stage moment selection procedures.

We provide power results that show that our test detects alternative parameter values that approach the boundary of the identified set at the fastest possible rate. The test is adaptive in the sense that it achieves these rates for data generating processes with a range of smoothness properties without prior knowledge of these smoothness properties. The test achieves these optimal rates adaptively even without the use of first stage moment selection procedures, and our results show that moment selection procedures have little or no first order effect on power in many settings. While moment selection procedures will have some effect in finite samples, the results suggest that our test is less sensitive to moment selection than many of the procedures available in the literature. This is a particularly positive result for researchers who prefer not to use pre-tests because of the introduction of arbitrary user

driven parameters or because of concerns about robustness. The test achieves rate optimal power adaptively without the need for such pre-tests, and the researcher need not worry when using this form of our procedure that performing such a pre-test would have had a dramatic effect on power.

The test statistic we consider presents several technical obstacles in deriving the asymptotic distribution. Because of the variance weighting, which is needed for our test to have good power properties, the test statistic takes a supremum over a sequence of random processes for which functional central limit theorems do not hold. While similar technical issues have been solved in other settings using approximations by sequences of gaussian processes (see, for example Bickel and Rosenblatt, 1973; Chernozhukov, Lee, and Rosen, 2009), the multiscale nature of our test statistic (as opposed to test statistics based on kernels with a fixed sequence of bandwidths), makes the rate of approximation too poor for our purposes. In addition, the test statistic we consider takes the supremum over a process that is nonstationary in ways that the previous literature has not dealt with, so even deriving the asymptotic distribution of the supremum of the approximating gaussian process would require new techniques.

To overcome this, we use methods for tail approximations to nonstationary, nongaussian processes, applying them directly to the process in the sample. We use methods from Chan and Lai (2006) to derive tail approximations directly using a combination of moderate deviations results and tail equicontinuity conditions, thereby circumventing the need for strong approximations. We verify these conditions for our test statistic directly, and use these results in the derivation of the extreme value distribution. While verifying these conditions can be challenging, we anticipate that the techniques introduced here will be useful in other problems in econometrics where intermediate strong approximations are not available or do not give the best results.

## 1.1 Related Literature

This paper is related to the literature on partial identification and, in particular, the literature on conditional moment inequalities. The tests proposed in this paper are most closely related to those studied by Armstrong (2011b) and Chetverikov (2012) (the results in the present paper were developed independently and around the same time as the latter paper). Armstrong (2011b) considers estimation of the identified set using conservative confidence regions. While those results could be used for the problem considered here, the methods of proof used in that paper lead to extremely conservative critical values that are too large to

be useful in most practical settings. Chetverikov (2012) uses a different form of a statistic similar to ours (the supremum is taken only over a finite set of bandwidths and points that cannot grow too quickly) and different methods of proof that avoid deriving an asymptotic distribution or even showing that one exists. From a practical perspective, our method delivers an analytic formula that can be used to compute a critical value that does not require simulation, and also proves the asymptotic validity of modified bootstrap procedures, while the approach taken in Chetverikov (2012) only allows for the latter result.

Papers proposing other approaches to inference on conditional moment inequalities include Andrews and Shi (2009), Kim (2008), Khan and Tamer (2009), Chernozhukov, Lee, and Rosen (2009), Lee, Song, and Whang (2011), Ponomareva (2010), Menzel (2008) and Armstrong (2011a). While these approaches are useful in many settings (for example, settings where point identification is likely, or where the researcher has prior knowledge of certain smoothness properties of the data generating process), they do not achieve optimal power adaptively in the generic set identified case considered here (see Armstrong, 2012, for local power results for some of these models). Indeed, the test statistic considered here can be thought of as introducing an optimal weighting to the statistics proposed by Andrews and Shi (2009) and Kim (2008), thereby allowing the tests to adaptively achieve optimal power in the set identified case, but leading to dramatically different behavior of the test statistics (and leading to the technical difficulties described above for deriving asymptotic distribution results). The tests considered in this paper can also be thought of as modifying the kernel based statistics of Chernozhukov, Lee, and Rosen (2009) and Ponomareva (2010) to a multiscale statistic that chooses the bandwidth automatically and adaptively. As discussed above, this also leads to difficult technical issues not encountered in the previous literature (the gaussian approximations used by those papers do not give good enough rates of approximation, and there is additional nonstationarity in the process since it is indexed by the bandwidth as well as the location), which the present paper uses new techniques to circumvent. In sum, none of the other approaches in the literature satisfy the optimality properties of adaptively achieving the best possible rate for detecting local alternatives in set identified models. This paper considers a test statistic that satisfies these optimality properties, and, because it differs in important ways from other statistics considered in the literature, requires new techniques to derive critical values and asymptotic distribution results.

This paper is also related to the broader literature on partial identification, including the problem of inference on finitely many unconditional moment inequalities. Articles that consider this problem include Andrews, Berry, and Jia (2004), Andrews and Jia (2008), Andrews

and Guggenberger (2009), Andrews and Soares (2010), Chernozhukov, Hong, and Tamer (2007), Romano and Shaikh (2010), Romano and Shaikh (2008), Bugni (2010), Beresteanu and Molinari (2008), Moon and Schorfheide (2009), Imbens and Manski (2004) and Stoye (2009). In addition, there have been a number of applications of partial identification, including the conditional moment inequality models considered here, going back at least to Manski (1990). There are too many references to name all of them here, but papers include Pakes, Porter, Ho, and Ishii (2006), Manski and Tamer (2002), and Ciliberto and Tamer (2009).

From a technical standpoint, this paper is related to other papers deriving extreme value results for supremum statistics. The literature goes back at least to Bickel and Rosenblatt (1973), and includes recent papers such as Chernozhukov, Lee, and Rosen (2009). The arguments used in the proof in this paper are substantially different, as they do not use intermediate approximations by gaussian processes. As discussed in more detail in Section 2, the multiscale nature of the test statistic considered here makes the rates in these approximations too poor for our purposes. Our result also differs in that the test statistic we consider takes a supremum over a process that is nonstationary in ways not considered in the previous literature. While extreme value results have been derived for nonstationary processes (see, for example, Lee, Linton, and Whang, 2009), these results use other aspects of the structure of these problems that do not apply in our case.

The test statistic considered in this paper is related to scan statistics considered in the statistics literature. This paper is also related to the literature on adaptive inference. In particular, Dumbgen and Spokoiny (2001) apply a similar approach to ours in a one dimensional gaussian setting. This paper contributes to these literatures by deriving extreme value approximations in a setting with a multidimensional, nongaussian, nonstationary process, which requires new techniques for the same reasons described above. Horowitz and Spokoiny (2001) propose a different test for a related goodness of fit testing problem. Those authors consider adaptivity with respect to a different class of alternatives than the one in this paper, leading to a different approach. In particular, Horowitz and Spokoiny (2001) consider minimax rates with respect to  $L_2$  distance in a two-sided testing problem. In the set identified parametric conditional moment inequality models considered in the present paper, euclidean distance on the parameter space translates to  $L_\infty$  distance for the conditional mean (see Armstrong, 2011b, and the power results in Section 4 of this paper), leading to a different approach (while we use a supremum statistic, Horowitz and Spokoiny use a supremum over integration based statistics).

## 1.2 Notation and Plan for Paper

We use the following notation throughout the rest of the paper. For observations  $\{Z_i\}_{i=1}^n$ , the sample mean of a function  $g$  is given by  $E_n g(Z_i) \equiv \frac{1}{n} \sum_{i=1}^n g(Z_i)$ . Inequalities are defined for vectors as holding elementwise. For vectors  $a$  and  $b$ ,  $a \wedge b$  is the elementwise minimum, and  $a \vee b$  is the elementwise maximum.

The rest of the paper is organized as follows. Section 2 describes the setup and gives the main asymptotic distribution result. Section 3 derives critical values for the test based on this result. Section 4 provides results on the power of the test. Section 5 reports the results of a monte carlo study. Section 6 reports the results of an illustrative empirical application. Section 7 concludes. Appendices to the main text contain proofs of the results in the main text, as well as some additional results mentioned in the main text, including versions of some of the results from the body of the paper that incorporate uniformity in the underlying distribution.

## 2 Setup and Asymptotic Distribution

We observe iid data  $\{X_i, W_i\}_{i=1}^n$  where  $X_i \in \mathbb{R}^{d_x}$  and  $W_i \in \mathbb{R}^{d_w}$ . We wish to test the null hypothesis

$$E(m(W_i, \theta)|X_i) \geq 0 \quad \text{a.s.} \quad (1)$$

where  $m : \mathbb{R}^{d_w} \times \Theta \rightarrow \mathbb{R}^{d_y}$  is a known measurable function and  $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$  is a fixed parameter value. We use the notation  $\bar{m}(\theta, x)$  to denote a version of  $E(m(W_i, \theta)|X_i = x)$ . Typically, the null (1) is tested for each value of  $\theta$  in order to obtain a confidence region for parameters that are consistent with the model. The model may not be point identified, in the sense that there may be more than one value of  $\theta$  consistent with (1), and the tests in this paper are specifically geared towards this case. In general, we denote by  $\Theta_0$  the identified set of parameter values that are consistent with the restrictions in (1):

$$\Theta_0 \equiv \{\theta \in \Theta | E(m(W_i, \theta)|X_i) \geq 0 \text{ a.s.}\}.$$

While the above setup considers only a single probability distribution, this is only for notational convenience. We show in Appendix A that our test controls the asymptotic size uniformly over appropriate classes of underlying distributions.

We note that, while the above setup is written in terms of a parametric model  $m(W_i, \theta)$ , our methods apply more generally to test the inequality  $E(Y_i|X_i) \geq 0$  a.s., where  $Y_i$  is any random variable satisfying certain regularity conditions below. The reason we impose this additional structure is that our tests are designed to have good power properties for values of  $\theta$  that violate the null, but are near the identified set  $\Theta_0$  of parameters that satisfy the null. Since our goal is to distinguish parameter values in  $\Theta_0$  from nearby parameter values outside of  $\Theta_0$ , we state our power results in terms of sequences of parameter values and the rate at which they approach the boundary of  $\Theta_0$  (see Section 4). By deriving our results in terms of alternative parameter values rather than mathematical notions of distances of data generating processes, we obtain power results that are immediately applicable to assessing the statistical accuracy of confidence regions based on our tests in economic models.

Consider the test statistic  $T_n = (T_{n,1}, \dots, T_{n,d_Y})$  where

$$T_{n,j} = T_{n,j}(\theta) \equiv \left| \inf_{I(s,t) \subseteq \hat{\mathcal{X}}, t \geq t_n} E_n \frac{m_j(W_i, \theta) I(s < X_i < s + t)}{\hat{\sigma}_{n,j}(s, t, \theta)} \right|_{-},$$

$t_n$  is a sequence going to zero,  $\hat{\mathcal{X}}$  is the convex hull of  $\{X_i\}_{i=1}^n$ ,  $I(s, t) = [s_1, s_1 + t_1) \times \dots \times [s_{d_X}, s_{d_X} + t_{d_X})$  and

$$\hat{\sigma}_{n,j}^2(s, t, \theta) \equiv E_n m_j(W_i, \theta)^2 I(s < X_i < s + t) - [E_n m_j(W_i, \theta) I(s < X_i < s + t)]^2.$$

We can form a test by rejecting for large values of  $S_n = S(T_n)$ , where  $S : \mathbb{R}^{d_Y} \rightarrow \mathbb{R}$  is some function that is nondecreasing in each argument. For concreteness, we take  $S$  to be function that takes the maximum of the components of  $T_n$ :

$$S_n = S_n(\theta) = \max_{1 \leq j \leq d_Y} T_{n,j}(\theta).$$

It is worth commenting on the properties of this test statistic that differ from other statistics for this problem, and how they lead to optimal power properties for set identified models. We discuss this briefly here, and refer the reader to Armstrong (2012), which contains a comprehensive treatment of the power properties of different approaches, for details. In testing  $E(m(W_i, \theta)|X_i) \geq 0$  a.s., one can use essentially any test statistic that estimates  $E(m(W_i, \theta)|X_i)$  and takes some function of this that is large in magnitude when this estimate is negative for some value of  $x$ . Most conditional mean estimates can be thought of as using an instrumental variables approach, where the inequality  $E(m(W_i, \theta)|X_i) \geq 0$  a.s. is transformed into a set of inequalities  $E m(W_i, \theta) g(X_i) \geq 0$  all where  $g$  ranges over a set  $\mathcal{G}_n$

that is infinite or increases with the sample size (e.g., a kernel estimator does this with the functions  $g$  given by  $h((X_i - x)/h_n)$  where  $h_n$  goes to zero at some rate and  $x$  ranges over the support of  $X_i$ ) and the inequality may only hold approximately if  $g$  is not positive everywhere (e.g. if higher order kernels or sieves are used). Once a class  $\mathcal{G}_n$  is decided on, one faces the decision of how to transform estimates of  $Em(W_i, \theta)g(X_i)$  into a statistic that is positive and large in magnitude whenever one of these estimates is negative and large in magnitude. This includes deciding on how to weight each function  $g$ , and how to combine them. For the latter problem, one can take some power of the negative part of the test statistic and add or integrate these over  $g$  (a Cramer-von Mises or CvM style approach), or take the maximum or supremum of the negative part (a Kolmogorov-Smirnov or KS approach). In addition, since the null space is composite, one faces a choice in how to pick the critical value, and, in particular, whether to choose a critical value based on the least favorable distribution in the null space where  $E(m(W_i, \theta)|X_i = x) = 0$  for all  $x$ , or whether to use a pre-testing procedure that determines where the equality may hold and uses smaller critical values based on the results of this procedure.

In sum, one faces the decision of (1) which instruments (or kernels or sieves, etc.) to use, (2) how to weight them, (3) how to combine them (integration or summing, or taking the supremum) and (4) how to choose the critical value. For (1), our test statistic uses a class of product kernels with all possible bandwidths. Using a class of functions with multiple scales, rather than a kernel function with a single bandwidth, allows the test to find the optimal bandwidth adaptively for a range of smoothness conditions. For (2), the test statistic  $S_n$  weights each function by its variance. This weighting is essential in allowing the test statistic to find the instrument function that balances bias and variance in an optimal way for detecting a given alternative, and the improvement in power in the set identified case can be thought of as an optimal weighting result for moment inequality models.

For (3) our test statistic uses a supremum (KS) criterion rather than a criterion based on sums or integrals (a CvM criterion). To understand why a KS approach leads to more power than a CvM approach, it is helpful to consider the relationship between the nonsimilarity of these tests on the boundary of the identified set and power at nearby alternatives. If a test statistic behaves differently depending on where  $\bar{m}(x, \theta) = 0$ , then using the most conservative critical value will lead to poor power in cases where nearby parameter values in the null space lead to the inequality binding on a small set. While moment selection procedures can help alleviate this, the procedures that are known to be robust in these settings are not aggressive enough to attain good power. KS statistics are less sensitive



to which moments bind since the supremum of  $k$  sample means increases at a  $\sqrt{\log k}$  rate, while the sum of the positive part increases at a polynomial rate in  $k$ . Thus, by using a KS criterion, our test statistic achieves good power without requiring moment selection procedures, and the power of the test is less sensitive to these procedures, so that the decision (4) has less impact on the power of the test.

We impose the following conditions.

**Assumption 2.1.**

a.) *The distribution of  $m(W_i, \theta)$  conditional on  $X_i$  satisfies the following conditions.*

i.) *There exists a  $\lambda > 0$  and a constant  $M_\lambda$  such that*

$$E(\exp(\lambda|m_j(W_i, \theta)|)|X_i) < M_\lambda \text{ a.s. all } 1 \leq j \leq d_Y.$$

ii.)  *$\text{var}(m_j(W_i, \theta)|X_i = x)$  is positive and continuous in  $x$  for all  $j$ .*

iii.)  *$\text{corr}(m_j(W_i, \theta), m_k(W_i, \theta)|X_i = x)$  is bounded away from 1 for all  $j \neq k$ .*

b.) *The support  $\mathcal{X}$  of  $X_i$  is a compact, convex Jordan measurable set with strictly positive measure, and  $X_i$  has a density  $f$  that is bounded away from zero on  $\mathcal{X}$ .*

c.)  *$t_n \rightarrow 0$  and  $nt_n^{d_X}/|\log t_n|^4 \rightarrow \infty$ .*

Part (a) imposes regularity conditions on the moments of  $m(W_i, \theta)$ . It is worth noting that, while we impose some mild smoothness assumptions on the conditional variance, we place no assumptions on the smoothness of the conditional mean. Thus, while the power of our test depends on the smoothness properties of the conditional mean, our test is robust to very nonsmooth data generating processes. The convexity assumption in part (b) is imposed to simplify certain parts of the proof, and could be relaxed. The condition on  $t_n$  in part (c) is, up to the  $|\log t_n|$  term, the best possible rate. As discussed further in Section D, other methods of deriving critical values for this test statistic would not allow  $t_n$  to decrease quickly enough for the statistic to have good power.

The following theorem gives the asymptotic distribution of this test statistic, and provides feasible critical values that can be calculated analytically. For a version of this theorem that incorporates uniformity in the underlying distribution, we refer the reader to Appendix A.

**Theorem 2.1.** *Suppose that the null hypothesis (1) and Assumption 2.1 hold for  $\theta$ . Let  $\hat{c}_n = \text{vol}(\hat{\mathcal{X}})/t_n^{d_X}$  and let  $a(\hat{c}_n) = (2n \log \hat{c}_n)^{1/2}$  and  $b(\hat{c}_n) = 2 \log \hat{c}_n + (2d_X - 1/2) \log \log \hat{c}_n -$*

$\log(2\sqrt{\pi})$ . Then, for any vector  $r \in \mathbb{R}^{d_Y}$ ,

$$\liminf_n P(a(\hat{c}_n)T_n - b(\hat{c}_n) \leq r) \geq P(Z \leq r)$$

where  $Z$  is a  $d_Y$  dimensional vector of independent standard type I extreme value random variables. If, in addition  $\bar{m}_j(\theta, x) = 0$  for all  $x$  and  $j$ , then

$$a(\hat{c}_n)T_n - b(\hat{c}_n) \xrightarrow{d} Z.$$

### 3 Inference

An immediate consequence of Theorem 2.1 is a method for choosing feasible critical values for the test statistic  $S_n(\theta)$  that can be computed analytically. By Theorem 2.1,  $a(\hat{c}_n)S_n - b(\hat{c}_n)$  is asymptotically bounded by a random variable that is the maximum of  $d_Y$  standard type I extreme value random variables. By the properties of extreme value random variables, this distribution is itself type I extreme value, with cdf  $\exp(-d_Y \exp(-r))$ . Some calculation leads to the rejection rule

$$\text{reject if } S_n(\theta) > \hat{q}_{1-\alpha} \quad \text{where } \hat{q}_{1-\alpha} \equiv \frac{\log(d_Y) - \log(-\log(1 - \alpha)) + b(\hat{c}_n)}{a(\hat{c}_n)}. \quad (2)$$

It follows from Theorem 2.1 that this test is asymptotically level  $\alpha$ . We record this result in the following theorem.

**Theorem 3.1.** *Suppose that the null hypothesis (1) holds for  $\theta$  and that Assumption 2.1 holds. Let  $\hat{q}_{1-\alpha}$  be as defined in (2). Then*

$$\limsup_n P(S_n(\theta) > \hat{q}_{1-\alpha}) \leq \alpha.$$

If, in addition,  $\bar{m}(\theta, x) = 0$  for all  $x \in \mathcal{X}$ , then

$$P(S_n(\theta) > \hat{q}_{1-\alpha}) \rightarrow \alpha.$$

#### 3.1 Simulated Critical Values

While the critical value given in (2) gives a valid asymptotically level  $\alpha$  test, this critical value is based on extreme value approximations that may perform poorly in finite samples in

certain situations. While our monte carlos suggest that the analytic critical values perform well in many cases encountered in practice, we also propose bootstrap or simulation based critical values to remedy these issues in cases where the extreme value approximations do poorly. The asymptotic validity of these critical values follows from the same results as the asymptotic distribution result in Theorem 2.1.

We define our simulated critical values as follows. For each  $j$ , let  $\hat{M}_n(x)$  be any random sequence of functions that take values in  $\mathcal{X}$  to  $d_Y \times d_Y$  symmetric, positive definite, matrices. We require that sequence of variance matrices given by  $\hat{M}_n(x)$  be continuous in  $x$  and have correlation coefficients bounded away from one uniformly over  $n$  with probability one. One can choose  $\hat{M}_n(x)$  to be an estimate of the conditional variance matrix of the  $m(W_i, \theta)$ , but this is not necessary, and  $\hat{M}_n(x)$  can even be chosen to be the constant function that takes all values to the identity matrix. For each repetition  $b$  of  $B$  simulations, we draw  $n$  independent outcome variables  $\{Y_i^{*,b}\}_{i=1}^n$  with  $Y_i^{*,b} \sim N(0, \hat{M}_n(X_i))$  independent across  $i$  and  $b$  conditional on the data. We form the test statistic  $S_{n,b}^*$  for this repetition by replacing  $m(W_i, \theta)$  with  $Y_i^{*,b}$  in the definition of the test statistic. The simulated critical value is given by the  $1 - \alpha$  quantile of this bootstrap distribution:

$$\hat{q}_{1-\alpha, \text{sim}} = \inf \left\{ r \left| \frac{1}{B} \sum_{b=1}^B I(S_{n,b}^* \leq r) \geq 1 - \alpha \right. \right\}. \quad (3)$$

The asymptotic validity of this test follows immediately from the version of Theorem 2.1 in Appendix A that incorporates uniformity in the underlying distribution.

**Theorem 3.2.** *Suppose that the null hypothesis (1) holds for  $\theta$  and that Assumption 2.1 holds. Let  $\hat{q}_{1-\alpha, \text{sim}}$  be as defined in (3). Then*

$$\limsup_n P(S_n(\theta) > \hat{q}_{1-\alpha, \text{sim}}) \leq \alpha.$$

*If, in addition,  $\bar{m}(\theta, x) = 0$  for all  $x \in \mathcal{X}$ , then*

$$P(S_n(\theta) > \hat{q}_{1-\alpha, \text{sim}}) \rightarrow \alpha.$$

### 3.2 Moment Selection Procedures and the Choice of $t_n$

The rejection probabilities of the tests defined above will converge to  $\alpha$  when the conditional mean  $\bar{m}_j(\theta, x)$  is equal to zero for all  $x \in \mathcal{X}$  for all  $j$ . If these inequalities only bind on a

subset of  $\mathcal{X}$ , the rejection probability will be strictly less than  $\alpha$ , and it would seem that there would be the potential for large power improvements at nearby alternatives by using a smaller critical value that take this into account. Perhaps surprisingly, it turns out that there will typically be no first order power improvement from doing this in our setting. While this result should certainly not be taken to mean that there will be no effect on power in finite samples, the result suggests that our procedure will be less sensitive to moment selection than other procedures in the literature for which moment selection has a large effect on power asymptotically (see, for example Armstrong, 2011a; Andrews and Shi, 2009).

To see why this holds, first, note that, it can be shown that, if  $\bar{m}_j(\theta, x) = 0$  for some  $j$  only for values in some subset  $\tilde{\mathcal{X}}$ , Theorem 2.1 will hold with  $\tilde{\mathcal{X}}$  replacing  $\mathcal{X}$ . Thus, if we use prior knowledge of such a set  $\tilde{\mathcal{X}}$  with strictly positive volume, or find such a set with a first stage test, we would obtain a critical value  $\hat{q}_{1-\alpha}$  with  $\mathcal{X}$  replaced by  $\tilde{\mathcal{X}}$ . But note that, regardless of  $\mathcal{X}$ , the critical value  $\hat{q}_{1-\alpha}$  satisfies

$$\hat{q}_{1-\alpha} \sim b(\hat{c}_n)/a(\hat{c}_n) \sim (2 \log \hat{c}_n)^{1/2}/n^{1/2} \sim [2 \log t_n^{-d_X} + \log \text{vol}(\mathcal{X})]^{1/2}/n^{1/2} \sim (2 \log t_n^{-d_X})^{1/2}/n^{1/2}$$

(and the same holds for the simulated critical value  $\hat{q}_{1-\alpha, \text{sim}}$ ). Thus, even with prior knowledge of the contact set, the contact set would have only a second order effect on the critical value.

The above calculations can also be used to understand the effect of the choice of the minimal window width  $t_n$  on the power of the test. Suppose that  $t_n$  is chosen proportional to  $n^{-\delta}$  for some  $0 < \delta < 1$ . Then, by the above calculations, we will have

$$\hat{q}_{1-\alpha} \sim (2d_X \delta \log n)^{1/2}/n^{1/2}.$$

As shown in Section 4, larger values of  $\delta$  are required to obtain optimal power properties for less smooth conditional means. While choosing a larger value of  $\delta$  does not affect the rate at which local alternatives can approach the null space and be detected (the test is adaptive with  $t_n$  decreasing as quickly as allowed), it does have a non negligible effect on power through larger critical values. If  $t_n$  is chosen as  $n^{-\delta_2}$  for some value  $\delta_2$  instead of some other value  $\delta_1$  where  $\delta_1 > \delta_2$ , the critical value will increase by a factor of  $(\delta_1/\delta_2)^{1/2}$ .

Note also that the critical value is, up to first order, the same as the critical value for a test that only takes the infimum over all  $s$  with  $t$  fixed at  $t_n$ , which would correspond to the kernel approach considered in Chernozhukov, Lee, and Rosen (2009) and Ponomareva (2010). Thus, in typical settings, there is no first order loss in power from considering larger

bandwidths using the multiscale approach in this paper even if the optimal bandwidth is known.

## 4 Local Power

This section derives asymptotic approximations to power functions by considering the power of these tests under sequences of alternative parameter values that approach the boundary of the identified set. Consider a parameter value  $\theta_0$  on the boundary of the identified set, and a sequence of local alternatives given by  $\theta_n = \theta_0 + ar_n$  for some vector  $a \in \mathbb{R}^{d_\theta}$  and some sequence of scalars  $r_n \rightarrow 0$ . We impose the following conditions (see Armstrong, 2011b, for verification in several examples of a set of conditions that imply Assumption 4.1).

### Assumption 4.1.

- a.)  $\bar{m}(\theta, x)$  is differentiable in  $\theta$  with derivative  $\bar{m}_\theta(\theta, x)$  that is continuous as a function of  $\theta$  uniformly in  $(\theta, x)$ .
- b.) For some  $\gamma, C, j$  and  $x_0 \in \mathcal{X}$ , we have  $\bar{m}_j(\theta_0, x_0) = 0$  and, for all  $x$  in a neighborhood of  $x_0$ ,

$$|\bar{m}_j(\theta_0, x) - \bar{m}_j(\theta_0, x_0)| \leq C\|x - x_0\|^\gamma.$$

Part (b) of Assumption 4.1 is a smoothness condition on the conditional mean under  $\theta_0$ . If  $\bar{m}_j(\theta_0, x_0) = 0$  for some  $x_0$ , part (b) will hold with  $\gamma = 1$  if  $\bar{m}_j(\theta_0, x)$  has a continuous first derivative in  $x$ , and it will hold with  $\gamma = 2$  if  $\bar{m}_j(\theta_0, x_0) = 0$  has a continuous second derivative in  $x$  and  $x_0$  is on the interior of  $\mathcal{X}$ .

The following theorem gives local power results for sequences of local alternatives. To state the results, let  $C(\cdot)$  be any bounded function on the unit sphere such that Assumption 4.1 holds with  $C$  replaced by  $C((x - x_0)/\|x - x_0\|)$ . We can always take this function to be a constant function under Assumption 4.1, but, using this notation, we can state power results that are more precise.

**Theorem 4.1.** *Suppose that Assumption 4.1 holds for  $\theta_0$  and that Assumption 2.1 holds with the constants in part (a) uniform over a neighborhood of  $\theta_0$ . Let  $\theta_n = \theta_0 + ar_n$  for some  $a \in \mathbb{R}^{d_\theta}$  and a sequence of scalars  $r_n \rightarrow 0$ . Suppose that, for some index  $j$  such that part (b)*

of Assumption 4.1 holds for  $j$ ,

$$\begin{aligned} & \liminf r_n \left( \frac{n}{2 \log t_n^{-d_X}} \right)^{\gamma/(d_X+2\gamma)} \\ & > - \left\{ \inf_{s,t} \frac{f(x_0)^{1/2} \int_{u \in \mathcal{U}, s < u < s+t} \left\{ [\bar{m}_{\theta,j}(\theta_0, x_0)a] + C \left( \frac{u}{\|u\|} \right) \|u\|^\gamma \right\} du}{\Sigma_{jj}^{1/2}(x) \text{vol}\{u \in \mathcal{U} | s < u < s+t\}^{1/2}} \right\}^{-\gamma/(d_X/2+\gamma)} \end{aligned}$$

where  $\mathcal{U} = \cup_{k=1}^{\infty} (\mathcal{X} - x_0)/r_k$ ,  $\Sigma_{jj}(x) = \text{var}(m_j(W_i, \theta) | X_i = x)$  and the right hand side is taken to be infinity if the infimum in the brackets is zero. Then, if  $t_n < \eta(n/\log n)^{-1/(d_X+2\gamma)}$  for small enough  $\eta$ , we will have

$$P(S_n(\theta_n) > \hat{q}_{1-\alpha}) \rightarrow 1.$$

If  $\bar{m}_{\theta_j}(\theta_0, x_0)a$  is strictly negative, which will typically be the case as long as  $\theta_n$  is outside of the identified set, this result shows that the power of the test approaches one as long as  $\theta_n$  approaches  $\theta_0$  at a  $(n/\log n)^{\gamma/(d_X+2\gamma)}$  rate with a large enough scaling. This corresponds to the fastest rate that would be achievable even if  $\gamma$  were known (see, for example, Stone, 1982). Theorem 4.1 shows that our test is adaptive in the sense that it achieves this rate simultaneously for all  $\gamma$  without prior knowledge of  $\gamma$ . Taking  $t_n$  to be a  $\log n$  term times  $n^{-1/d_X}$ , the condition that  $t_n < \eta(n/\log n)^{-1/(d_X+2\gamma)}$  will be satisfied regardless of  $\gamma$ . Another possibility is to take the smallest value of  $\gamma$  that the researcher thinks is likely, and to choose a value of  $t_n$  that is optimal for a particular data generating process and sequence of alternatives with that value of  $\gamma$ . Theorem 4.1 shows that this approach will achieve the optimal rate even if  $\gamma$  is larger than the value used to choose  $t_n$ .

## 5 Monte Carlo

We perform monte carlos with several designs based on a median regression model with potentially endogenously missing data. We consider a missing data model where the conditional median of  $W_i^*$  given  $X_i$  is given by  $q_{1/2}(W_i^* | X_i) = \theta_1 + \theta_2 X_i$ , and  $W_i^*$  is missing for some observations. Letting  $W_i^H = W_i^*$  when  $W_i^*$  is observed and  $\infty$  otherwise, this leads to the conditional moment inequality  $E[I(\theta_1 + \theta_2 X_i \leq W_i^H) - 1/2 | X_i] \geq 0$  a.s. (in practice, one would form another inequality based on a lower bound for  $W_i^*$  of  $-\infty$  when  $W_i^*$  is not observed, but we focus on a single moment inequality in the monte carlos for simplicity).

In each design, we simulate the data from a median regression given by  $W_i^* = \theta_1^* + \theta_2^* X_i + u$  for some  $(\theta_1^*, \theta_2^*)$  where  $u \sim \text{unif}(-1, 1)$  and  $X_i \sim \text{unif}(0, 1)$ . We then set  $W_i^*$  to be missing with probability  $p(X_i)$  independently of  $W_i^*$  for some function  $p(x)$  (note that, while we generate the data using a parameter value that satisfies missingness at random, the test is designed to give confidence regions that are robust to the failure of this assumption). We consider 3 designs with  $\theta_1^* = \theta_2^* = 0$  and  $p(x)$  given as follows for each design:

$$\begin{aligned} \text{Design 1: } & p(x) = .1 \\ \text{Design 2: } & p(x) = .02 + 2 \cdot .98 \cdot |x - .5| \\ \text{Design 3: } & p(x) = .02 + 4 \cdot .98 \cdot (x - .5)^2. \end{aligned}$$

Design 1 corresponds to a flat conditional mean, while Designs 2 and 3 correspond to  $\gamma = 1$  and  $\gamma = 2$  in Assumption 4.1 respectively. For each design, we consider the sample sizes  $n = 100, 500, 1000$  and the truncation parameters  $t_n = n^{-1/5}, n^{-1/3}, n^{-1/2}$  for each sample size. Note that  $n^{-1/3}$  is the optimal rate for  $t_n$  for Design 2 and  $n^{-1/5}$  is the optimal rate for  $t_n$  for Design 3, while  $t_n = n^{-1/2}$  is smaller than optimal for all three designs, but still achieves the optimal rate for local alternatives by Theorem 4.1.

For each design, we test several parameter values with  $\theta_2$  fixed at 0 and  $\theta_1$  varying. For a given design, let  $\bar{\theta}_1$  be the largest value of  $\theta_1$  such that  $(\theta_1, 0)$  is in the identified set. First, to examine the finite sample size of the test based directly on the asymptotic distribution, we report monte carlo estimates of the true false rejection probability under  $(\bar{\theta}_1, 0)$  and Design 1, which corresponds to a least favorable null distribution with the conditional moment inequality equal to zero for all  $x$ . This gives an idea of the worst (most liberal) size distortions one can expect from tests based on critical values calculated directly from the asymptotic distribution (at least, in situations similar to the median regressions with potentially endogenously missing or censored data considered here).

Table 1 reports these results. We note that size distortions are generally minimal, except for the smaller sample sizes with the largest value of  $t_n$ , particularly with nominal size  $\alpha = .1$ . As one might expect from the methods used in the derivation of the asymptotic distribution, which rely on tail approximations, the asymptotic approximation performs better for the smaller value of the nominal size  $\alpha$ . The fact that size distortions are more severe with the larger  $t_n = n^{-1/5}$  is likely a reflection of the fact, for a fixed nominal size  $\alpha$ , the asymptotic approximations depend on  $t_n$  being small relative to the support of  $X_i$ . In contrast, size distortions are minimal for  $t_n = n^{-1/3}$  for most cases considered here.

Next, we examine the power of our test. We report monte carlo estimates of the power

of our test for each design and parameters given by  $(\bar{\theta}_1 + a, 0)$  for  $a = .1, .2, .3, .4, .5$ . To ensure that power is not driven by false rejection under the null, we use critical values based on monte carlo estimates of the finite sample exact least favorable distribution. We report power results for level .05 tests. Tables 2, 3 and 4 report the results. As expected, moving away from the identified set by a given amount generally leads to more power under the designs with smoother conditional means. In addition, the finding that the choice of the truncation parameter  $t_n$  doesn't matter much as long as it is small enough appears to be borne out in the monte carlos (e.g. for Design 2,  $t_n$  proportional to  $n^{-1/3}$  is optimal, and this value of  $t_n$  performs best, but choosing  $t_n = n^{-1/2}$  gives close to the same power, while  $t_n = n^{-1/5}$  gives much worse power).

## 6 Empirical Illustration

We apply our methods to a median regression model with endogenously censored and missing data, using data from the Health and Retirement Study. The setup follows Section 9 of Armstrong (2011a), but we repeat it here for convenience. Letting  $X_i$  and  $W_i^*$  be yearly income and prescription drug expenditures for participant  $i$  respectively, we posit the model

$$q_{1/2}(W_i^*|X_i) = \theta_1 + \theta_2 X_i \quad (4)$$

where  $q_{1/2}(W_i^*|X_i)$  is the median of  $W_i^*$  conditional on  $X_i$ .

In this survey, participants who did not report a point value for prescription drug expenditures were given a series of brackets for this variable, resulting in interval censoring for a portion of the observations, and some observations with a completely missing outcome variable. In other words, we do not observe  $W_i^*$ , but only observe a random interval  $[W_i^L, W_i^H]$  known to contain  $W_i^*$ . The data is censored in a way that is likely to violate a missingness at random or censoring at random assumption: the variable is censored only for those who do not recall how much they spent, and it is likely that remembering how much one spent is correlated with the level of spending itself.

This endogenous censoring problem makes it impossible to estimate  $(\theta_1, \theta_2)$  consistently in general. We construct bounds using the conditional moment inequalities

$$E[m(X_i, W_i^L, W_i^H, \theta)|X_i] \equiv E \left[ \begin{array}{c} I(\theta_1 + \theta_2 X_i \leq W_i^H) - 1/2 \\ 1/2 - I(\theta_1 + \theta_2 X_i \leq W_i^L) \end{array} \middle| X_i \right] \geq 0 \quad \text{a.s.} \quad (5)$$



We test (5) at the .05 level using our methods for each value of  $(\theta_1, \theta_2)$ , and report a 95% confidence region that inverts these tests. The resulting confidence region contains the true parameter value with probability at least .95.

We restrict our sample to the 1996 wave of the survey and women with no more than \$15,000 of yearly income who report using prescription medications. The data set also contains observations with a censored covariate (income), but, for illustrative purposes, we focus on endogenous censoring of the outcome variable and throw away observations where income is missing or censored (this is valid if remembering prescription drug expenditures is not correlated with income, but may be correlated with spending itself). Our data set has 636 observations, of which 54 have an interval censored outcome variable, and an additional 7 have a completely missing outcome variable. See Armstrong (2011a) for additional details about the data set. For the truncation parameter  $t_n$ , we use  $n^{-1/3} \cdot (\max_{1 \leq i \leq n} X_i - \min_{1 \leq i \leq n} X_i)$ . The  $n^{-1/3}$  scaling results in a test statistic that is rate adaptive to smoothness between Lipschitz continuity and 2 derivatives of the conditional truncation probabilities (a smaller value could be used to adapt to a less smooth data generating process). For the critical value for our test, we use the analytically computed critical value defined in (2).

Figure 1 shows the resulting confidence region. For comparison, Figures 2 and 3 show confidence regions using the tests proposed in Armstrong (2011a) and Andrews and Shi (2009) respectively, taken directly from Armstrong (2011a). The test considered in this paper can be thought of as introducing an optimal weighting to the Andrews and Shi (2009) statistic that improves the rate for local alternatives from  $n^{-\gamma/(2d_x+2\gamma)}$  to the  $(n/\log n)^{\gamma/(d_x+2\gamma)}$  rate obtained in Theorem 4.1 in the set identified case, while reducing the rate by a  $\log n$  term in the point identified case. The Armstrong (2011a) test yields a slightly better improvement in power, but is not robust to failure of certain smoothness conditions. We also report confidence regions for each component of  $(\theta_1, \theta_2)$ , formed by projecting the confidence region onto each component. Table 5 reports these confidence intervals, along with the corresponding confidence intervals formed using other methods reproduced from Armstrong (2011a) for convenience.

The slope parameter,  $\theta_2$ , gives the median increase in yearly prescription drug spending associated with an increase in income. Thus, according to the results using the test proposed in this paper, a 95% confidence interval puts the median increase in prescription drug expenditures associated with a \$1,000 in income between \$5.30 and \$32.00. It is worth making a few notes in comparing this with the confidence regions using the unweighted statistic. As predicted by the asymptotic power results, the confidence region for the slope parameter is

tighter than the one obtained using an unweighted test statistic with a critical value formed using subsampling with a conservative rate. The unweighted statistic gives a better lower bound for the slope parameter when subsampling with an estimated rate is used to form the critical value, but this test is less robust in the sense that it relies on additional smoothness conditions.

Comparing the joint confidence regions for  $(\theta_1, \theta_2)$ , we see that the tests based on unweighted statistics with subsampling based critical values lead to disconnected regions of rejected and accepted parameter values. While the test based on a conservative rate proposed in Andrews and Shi (2009) has only a small island of rejected parameter values in the confidence region, the test based on an estimated rate proposed in Armstrong (2011a) leads to numerous isolated areas in the confidence region. In contrast, our test leads to a connected confidence region. A likely explanation for this phenomenon is that the subsampling based confidence regions use critical values that implicitly estimate where the data generating process is in the null space. This leads to disconnected confidence regions when, as the parameter moves in some direction, the test first begins to reject as the test statistic increases, but then fails to reject when the critical value increases as well. In contrast, our test uses a least favorable critical value, so the test always moves from acceptance to rejection as the test statistic increases.

## 7 Conclusion

This paper considers inference in conditional moment inequality models using a multiscale statistic. The asymptotic distribution of our test statistic is derived, and the results are used to obtain feasible critical values. The test is shown to obtain certain optimal rates for power against local alternatives adaptively, and is the only feasible test available that does so for the best possible range of smoothness classes. Our results also have implications for the effect of moment selection procedures on power, and our test has the additional advantage of being adaptive without requiring such tests. An empirical application to a regression model with endogenous censoring and missing data illustrates the power improvement from the test.

# A Uniformity in the Underlying Distribution

We prove a stronger version of Theorem 2.1 that holds uniformly in certain classes  $\mathcal{P}$  of underlying distributions for which Assumption 2.1 holds uniformly over  $P \in \mathcal{P}$ . To state and prove this result, we introduce some notation for indexing certain quantities by the underlying distribution  $P$ . We use the notation  $E_P$  to denote expectation with respect to the probability distribution  $P$ , and use similar notation for conditional expectations and conditional and unconditional variances, covariances and correlations. We make explicit the dependence of the identified set on  $P$  and define  $\Theta_0(P) = \{\theta \in \Theta | E_P[m(W_i, \theta) | X_i] \geq 0 \text{ a.s.}\}$ .

In the following theorem, the conditional distribution (including the conditional mean) of  $m(W_i, \theta)$  given  $X_i = x$  is allowed to vary over  $\mathcal{P}$ . In particular, since no conditions are placed on the conditional mean of distributions in  $\mathcal{P}$ , the result shows that tests based on this asymptotic distribution result control the asymptotic size uniformly over distributions for which the conditional mean can be nonsmooth in arbitrary ways, although there are some mild continuity assumptions on the conditional variance. We do, however, impose the same distribution of  $X_i$  for all  $P \in \mathcal{P}$ . This is mostly to avoid introducing additional notation in the proof, and could be relaxed.

**Theorem A.1.** *Let  $\hat{c}_n$ ,  $a(\hat{c}_n)$  and  $b(\hat{c}_n)$  be defined as in Theorem 2.1. Suppose that Assumption 2.1 holds for the same constants in part (a) for all  $P \in \mathcal{P}$  and with the continuity in part (ii) of part (a) uniform over  $P \in \mathcal{P}$ . Then, for any vector  $r \in \mathbb{R}^{d_Y}$ ,*

$$\liminf_n \inf_{P \in \mathcal{P}, \theta_0 \in \Theta_0(P)} P(a(\hat{c}_n)T_n(\theta_0) - b(\hat{c}_n) \leq r) \geq P(Z \leq r)$$

where  $Z$  is a  $d_Y$  dimensional vector of independent standard type I extreme value random variables. If, in addition,  $E_P[m(W_i, \theta_0) | X_i] = 0$  for all  $P \in \mathcal{P}$  for some  $\theta_0$ , then, for this  $\theta_0$ ,

$$a(\hat{c}_n)T_n - b(\hat{c}_n) \xrightarrow{d} Z$$

uniformly over  $P \in \mathcal{P}$ .

For completeness, we also include the following theorem, which states that the tests proposed in this paper control the size uniformly over classes of distributions that satisfy the conditions of the above theorem.

**Theorem A.2.** *For any class  $\mathcal{P}$  of distributions satisfying the conditions of Theorem A.1,*

$$\limsup_n \sup_{P \in \mathcal{P}, \theta_0 \in \Theta_0(P)} P(S_n(\theta_0) > \hat{q}_{1-\alpha}) \leq \alpha,$$

*and the same holds with  $\hat{q}_{1-\alpha}$  replaced by  $\hat{q}_{1-\alpha, sim}$  as long as the sequence of bootstrap conditional variance matrices is continuous uniformly over  $x$  and has correlation coefficients bounded away from one with probability one with the continuity and bounds on the correlation coefficients uniform over  $P \in \mathcal{P}$ .*

Theorem A.2 follows immediately from Theorem A.1 (for the validity of the critical value  $\hat{q}_{1-\alpha, sim}$ , the result follows from applying Theorem A.1 to the sequence of bootstrap processes). We prove Theorem A.1 in the next appendix.

## B Proof of Theorem A.1

We first prove a version of Theorem A.1 where the  $X_i$ s are deterministic and  $\hat{\sigma}^2$  is replaced by a certain sample average of conditional variances defined below. The result then follows from showing that the conditions of this result hold almost surely conditional on  $\{X_i\}_{i=1}^n$ , and that replacing the sample average of conditional variances with  $\hat{\sigma}^2$  does not change the test statistic too much.

Throughout this section, we fix  $\theta$  and let  $Y_i = m(W_i, \theta)$ , and drop the  $\theta$  notation elsewhere such as in the definition of  $\hat{\sigma}_{n,j}(s, t, \theta)$ . We prove the following result with  $\{X_i\}_{i=1}^n$  replaced by a deterministic sequence  $\{x_i\}_{i=1}^n$ . We consider a set  $\mathcal{P}$  determining the probability distribution of  $Y_i$  for a given  $x_i$ .

Let  $\mathcal{F} = \{F_{x,P} : x \in \mathcal{X}, P \in \mathcal{P}\}$  be a family of  $d_Y$ -dimensional distribution functions, with  $\mathcal{X}$  a compact, Jordan measurable subset of  $\mathbb{R}^{d_X}$  such that  $\text{vol}(\mathcal{X}) > 0$ , that is it has positive  $d_X$  dimensional volume. Consider  $(x_1, Y_1), (x_2, Y_2), \dots$  with  $x_i$  deterministic and  $Y_i \sim F_{x_i}$  independent. Define  $\mu_P(x) = E_{x,P} Y_i$  and  $\Sigma_P(x) = \text{Cov}_{x,P} Y_i$ , where the subscript  $x, P$  denotes with respect to  $Y_i \sim F_{x,P}$ . We use the notation  $z_{i,j}$  to denote the  $j$ th coordinate of the  $i$ th observation or element in a sequence  $\{z_i\}$ . Let  $I(s, t) = \prod_{j=1}^{d_t} [s_j, s_j + t_j]$ . We abuse notation slightly and define  $\text{vol}(t) = \prod_{j=1}^{d_t} t_j$  for a vector  $t$ . Let  $J_n(s, t) = \{i : 1 \leq i \leq n, x_i \in I(s, t)\}$ . We consider the following regularity conditions.

**Assumption B.1.**

a.) There exists  $\lambda > 0$  and  $M_\lambda < \infty$  such that

$$E_{x,P}(e^{\lambda|Y_{i,j}|}) \leq M_\lambda \text{ for all } x \in \mathcal{X}, 1 \leq j \leq d_Y, P \in \mathcal{P}.$$

Hence the characteristic function of  $Y_{i,j}$  is analytic on  $(-\lambda, \lambda)$  for all  $j$  and  $Y_i \sim F_{x,P}$ ,  $x \in \mathcal{X}$ ,  $P \in \mathcal{P}$ .

b.)  $\sigma_{j,P}(x) \equiv \Sigma_{jj,P}^{1/2}(x)$  is continuous and positive on  $\mathcal{X}$  for all  $1 \leq j \leq d_Y$  uniformly over  $P \in \mathcal{P}$ .

c.) (for  $d_Y > 1$ ):

$$\rho \equiv \sup_{P \in \mathcal{P}} \sup_{i \neq j} \sup_{x \in \mathcal{X}} \frac{\Sigma_{ij,P}(x)}{\sigma_{i,P}(x)\sigma_{j,P}(x)} < 1.$$

**Assumption B.2.** There exists a continuous, positive and bounded density function  $f$  on  $\mathcal{X}$  and a sequence  $t_n \rightarrow 0$  such that

a.)  $nt_n^{d_X} |\log t_n|^{-4} \rightarrow \infty$ ,

b.) for any  $\delta > 0$ ,  $\#J_n(s, t) \sim n \int_{I(s,t)} f(x) dx$  uniformly over  $I(s, t) \subseteq \mathcal{X}$  such that  $\text{vol}(t) \geq \delta t_n^{d_X} / |\log t_n|^2$ .

Define  $\sigma_{n,j}(s, t) = \{\sum_{i \in J_n(s,t)} [\sigma_{j,P}(x)]^2\}^{1/2}$  and let

$$\tilde{T}_{n,j} = - \inf_{I(s,t) \subseteq \mathcal{X}, t \geq t_n} \sum_{i \in J_n(s,t)} Y_{i,j} / [\sqrt{n} \sigma_{n,j}(s, t)]$$

(we suppress the dependence of  $\sigma_{n,j}(s, t)$  and  $\tilde{T}_{n,j}$  on  $P$  for notational convenience).

**Theorem B.1.** Suppose that  $\mu_P(x) \geq 0$  for all  $x \in \mathcal{X}$ ,  $P \in \mathcal{P}$  and that Assumptions B.1 and B.2 hold. Let  $a_n = (2n \log t_n^{-d_X})^{1/2}$  and  $b_n = 2 \log t_n^{-d_X} + (2d_X - \frac{1}{2}) \log \log t_n^{-d_X} - \log[2\sqrt{\pi}/\text{vol}(\mathcal{X})]$ . Then, for any vector  $r \in \mathbb{R}^{d_Y}$ ,

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}} P \left( a_n \tilde{T}_n - b_n \leq r \right) \geq P(Z \leq r)$$

where  $Z$  is a  $d_Y$  dimensional vector of independent standard type I extreme value random variables. If, in addition,  $\mu_P(x) = 0$  for all  $x \in \mathcal{X}$ ,  $P \in \mathcal{P}$ , then

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \left| P \left( a_n \tilde{T}_n - b_n \leq r \right) - P(Z \leq r) \right| = 0.$$

The result follows from this and the following lemmas.

**Lemma B.1.** *Under Assumption 2.1, part (b) of Assumption B.2 above holds for almost all sequences  $\{X_i\}_{i=1}^n$ .*

*Proof.* We have

$$\frac{\#J_n(s, t)}{n \int_{I(s, t)} f(x) dx} - 1 = \frac{E_n I(s < X_i < s + t) - E I(s < X_i < s + t)}{E I(s < X_i < s + t)}.$$

This converges to one uniformly over  $(s, t)$  with  $\text{vol}(t) \geq K_n(\log n)/n$  for any sequence  $K_n \rightarrow \infty$  by Theorem 37 in Chapter 2 of Pollard (1984), and the conditions  $nt_n^{d_x}/|\log t_n|^4 \rightarrow \infty$  and  $\text{vol}(t) \geq \delta t_n^{d_x}/|\log t_n|^2$  guarantee that  $\text{vol}(t) \geq \delta t_n^{d_x}/|\log t_n|^2 \geq K_n n^{-1} |\log t_n|^4 / |\log t_n|^2 \geq K_n(\log n)/n$  for some  $K_n \rightarrow \infty$ .  $\square$

**Lemma B.2.** *Under Assumption 2.1,  $\text{vol}(\hat{\mathcal{X}}) \xrightarrow{P} \text{vol}(\mathcal{X})$ .*

*Proof.* For a given  $\varepsilon, \delta > 0$ , the following event will hold with probability approaching one: for every point  $\varepsilon k$  in the grid  $(\varepsilon \mathbb{Z}^{d_x}) \cap \mathcal{X}$ , at least one observation  $X_i$  will have each component  $X_{i,j}$  within  $\delta$  of  $\varepsilon k$ . Once this holds, the set  $\varepsilon I((k_1 + \delta, \dots, k_{d_x} + \delta), (1 - \delta, \dots, 1 - \delta))$  will be contained in the convex hull of the  $X_i$ s for all  $k$  such that  $\varepsilon I(k, \mathbf{1}) \subseteq \mathcal{X}$ . This gives a lower bound of  $(1 - 2\delta)^{d_x} \text{vol}(\cup_{\varepsilon I(k, \mathbf{1}) \subseteq \mathcal{X}} \varepsilon I(k, \mathbf{1}))$  for the volume of the convex hull of the  $X_i$ s, which can be made arbitrarily close to  $\text{vol}(\mathcal{X})$  by Jordan measurability. The result follows from this and the upper bound  $\text{vol}(\hat{\mathcal{X}}) \leq \text{vol}(\mathcal{X})$ .  $\square$

**Lemma B.3.** *Under Assumptions B.1 and B.2 (with the  $X_i$ 's treated as nonrandom),  $\sup_{s, s+t \in \hat{\mathcal{X}}, t \geq t_n} \frac{\sigma_{n,j}(s, t)}{\sqrt{n} \hat{\sigma}_{n,j}(s, t)} - 1 \leq o_P(\log n)^{-1}$  uniformly over  $P \in \mathcal{P}$  and, if  $\bar{m}(\theta, x) = 0$  for all  $x$ ,  $\sup_{s, s+t \in \hat{\mathcal{X}}, t \geq t_n} \left| \frac{\sigma_{n,j}(s, t)}{n \hat{\sigma}_{n,j}(s, t)} - 1 \right| = o_P(\log n)^{-1}$  uniformly over  $P \in \mathcal{P}$ .*

*Proof.* First, note that, since  $x \mapsto 1/x^2$  is decreasing and differentiable at one, it suffices to show that  $\inf_{s, s+t \in \hat{\mathcal{X}}} \frac{n \hat{\sigma}_{n,j}^2(s, t)}{\sigma_{n,j}^2(s, t)} - 1 \geq -o_P(\log n)^{-1}$  and  $\sup_{s, s+t \in \hat{\mathcal{X}}} \left| \frac{n \hat{\sigma}_{n,j}^2(s, t)}{\sigma_{n,j}^2(s, t)} - 1 \right| = o_P(\log n)^{-1}$ . Note that

$$\hat{\sigma}_{n,j}^2(s, t) - \sigma_{n,j}^2(s, t)/n = \frac{1}{n} \sum_{i \in J_n(s, t)} Y_{i,j}^2 - \left[ \frac{1}{n} \sum_{i \in J_n(s, t)} Y_{i,j} \right]^2 - \frac{1}{n} \sum_{i \in J_n(s, t)} \sigma_{j,P}(x)^2 = I + II$$

where

$$I \equiv \frac{1}{n} \sum_{i \in J_n(s, t)} (Y_{i,j}^2 - E_{x_i} Y_{i,j}^2)$$

and

$$II \equiv \frac{1}{n} \sum_{i \in J_n(s,t)} [E_{x_i} Y_{i,j}^2 - \sigma_{j,P}(x)^2] - \left[ \frac{1}{n} \sum_{i \in J_n(s,t)} Y_{i,j} \right]^2 = \frac{1}{n} \sum_{i \in J_n(s,t)} [E_{x_i} Y_{i,j}]^2 - \left[ \frac{1}{n} \sum_{i \in J_n(s,t)} Y_{i,j} \right]^2.$$

We first bound  $I/[\sigma_{n,j}^2(s,t)/n]$  where  $I$  is given above. Let  $W_i = Y_{i,j}^2 - E_{x_i,P} Y_{i,j}^2$ . Note that  $\sigma_{n,j}^2(s,t)$  is bounded from below by a constant times  $\#J_n(s,t)$  uniformly over  $P \in \mathcal{P}$ , so it suffices to consider  $(\sum_{i \in J_n(s,t)} W_i) / \#J_n(s,t)$ . For some sequence  $K_n$ , let  $\tilde{W}_i = W_i I(|W_i| \leq K_n)$  be a truncated version of  $W_i$ . Note that, by Markov's inequality, for  $\lambda > 0$  given in Assumption B.1,

$$P(|W_i| > K) \leq E_{x_i,P} \exp(\lambda \sqrt{|W_i|} - \lambda \sqrt{K})$$

so

$$P(|W_i| > K \text{ some } 1 \leq i \leq n) \leq n \exp(-\lambda \sqrt{K}) \sup_{x \in \mathcal{X}, P \in \mathcal{P}} E_{x,P} \exp(\lambda \sqrt{|W_i|}),$$

which goes to zero for any  $K = K_n$  that increases faster than  $(\log n)^2$ . To bound  $|E_{x_i,P} \tilde{W}_i| = |E_{x_i,P} \tilde{W}_i - E_{x_i,P} W_i|$ , note that

$$\{E_{x_i,P}[|W_i| I(|W_i| > K)]\}^2 \leq E_{x_i,P}(W_i^2) P(|W_i| > K) \leq C \exp(-\lambda \sqrt{K})$$

for some constant  $C$  that does not depend on  $P$  or  $x_i$ . Thus,  $|\sum_{i \in J_n(s,t)} E_{x_i,P} \tilde{W}_i| / \#J_n(s,t) \leq [C \exp(-\lambda \sqrt{K_n})]^{1/2}$ , which goes to zero at a polynomial rate for  $K_n$  increasing faster than  $(\log n)^2$ , which is faster than the required  $\log n$  rate.

Using the fact that the supremum over  $(s,t)$  is determined by the maximum over no more than  $n^{2d_x}$  possible deterministic configurations for  $J_n(s,t)$ , and that for any  $\delta > 0$ ,  $\delta(\log n)^{-1} \geq \#J_n(s,t)^{-1/4}$  for large enough  $n$ ,

$$\begin{aligned} & P \left( \sup_{s,s+t \in \hat{\mathcal{X}}, t \geq t_n} \left| \frac{\sum_{i \in J_n(s,t)} [\tilde{W}_i - E_{x_i,P} \tilde{W}_i]}{\#J_n(s,t)} \right| \geq \delta(\log n)^{-1} \right) \\ & \leq n^{2d_x} \sup_{s,s+t \in \hat{\mathcal{X}}, t \geq t_n} P \left( \left| \frac{\sum_{i \in J_n(s,t)} [\tilde{W}_i - E_{x_i,P} \tilde{W}_i]}{\#J_n(s,t)} \right| \geq \#J_n(s,t)^{-1/4} \right). \end{aligned}$$

Now, using Bernstein's inequality, for  $C$  a bound for the fourth moment of  $Y_{i,j}$ , the above

display is bounded by

$$n^{2d_X} \sup_{s, s+t \in \hat{\mathcal{X}}, t \geq t_n} 2 \exp \left( - \frac{[\#J_n(s, t)^{3/4}]^2}{C\#J_n(s, t) + K_n[\#J_n(s, t)^{3/4}]/3} \right).$$

Let  $K_n$  be such that  $K_n \leq \#J_n(s, t)^{1/2}$  all  $(s, t)$  and  $K_n/(\log n)^2 \rightarrow \infty$ . For large enough  $n$ , this gives a bound in the above display of  $n^{2d_X} \sup_{s, s+t \in \hat{\mathcal{X}}, t \geq t_n} \exp(-\#J_n(s, t)^{1/4}) \rightarrow 0$ .

As for  $II$ , we have

$$\begin{aligned} II &\geq \left[ \frac{1}{n} \sum_{i \in J_n(s, t)} E_{x_i, P} Y_{i, j} \right]^2 - \left[ \frac{1}{n} \sum_{i \in J_n(s, t)} Y_{i, j} \right]^2 \\ &\geq -2 \left( \left| \frac{1}{n} \sum_{i \in J_n(s, t)} E_{x_i, P} Y_{i, j} \right| \vee \left| \frac{1}{n} \sum_{i \in J_n(s, t)} Y_{i, j} \right| \right) \left| \frac{1}{n} \sum_{i \in J_n(s, t)} (Y_{i, j} - E_{x_i, P} Y_{i, j}) \right| \end{aligned}$$

and similar methods show that the last line divided by  $\sigma_{n, j}(s, t)/n$  converges to zero at a faster than  $\log n$  rate uniformly over  $(s, t)$  with  $s, s+t \in \hat{\mathcal{X}}, t \geq t_n$ . If  $\bar{m}_j(\theta, x) = 0$  for all  $x$ , then  $II = - \left[ \frac{1}{n} \sum_{i \in J_n(s, t)} Y_{i, j} \right]^2$ , and  $\left[ \frac{1}{n} \sum_{i \in J_n(s, t)} Y_{i, j} \right]^2 / [\sigma_{n, j}(s, t)/n]$  also converges to zero at a faster than  $\log n$  rate uniformly over  $(s, t)$  with  $s, s+t \in \hat{\mathcal{X}}, t \geq t_n$  by similar arguments.  $\square$

## B.1 Proof of Theorem B.1

We begin by proving the result in the case of a univariate outcome  $Y_i = m(W_i, \theta)$ . Section B.4 generalizes the result to the case of multivariate  $Y_i$ .

To simplify notation, we let  $d = d_X$  and we omit the subscript  $P$  when dealing with expectations and other quantities that depend on the underlying distribution  $P$ . Let  $Z_i = \mu(x_i) - Y_i$  and let  $B_n = \{(s, t) : I(s, t) \subseteq \mathcal{X}, t \geq t_n \mathbf{1}\}$ . Define

$$\mathbb{H}_n(s, t) = \frac{\sum_{i \in J_n(s, t)} Z_i}{\sigma_n(s, t)} \text{ for } (s, t) \in B_n. \quad (6)$$



Let  $c = \frac{(b_n + \zeta)\sqrt{n}}{a_n}$ . Note in particular that

$$\begin{aligned}
& t_n^{-d} (c^2/2)^{2d-\frac{1}{2}} e^{-c^2/2} \\
& \sim t_n^{-d} (d|\log t_n|)^{2d-\frac{1}{2}} \exp\left(-\frac{1}{2}\left\{(2d|\log t_n|)^{1/2} + \frac{\log[(d|\log t_n|)^{2d-\frac{1}{2}} \text{vol}(\mathcal{X})e^\zeta/2\sqrt{\pi}]}{(2d|\log t_n|)^{1/2}}\right\}^2\right) \\
& \rightarrow [2\sqrt{\pi}/\text{vol}(\mathcal{X})]e^{-\zeta} \text{ as } n \rightarrow \infty.
\end{aligned} \tag{7}$$

Theorem B.1 in the case of univariate  $Y$  follows from

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \left| P\left\{ \sup_{(s,t) \in B_n} \mathbb{H}_n(s,t) \geq c \right\} - [1 - \exp(-e^{-\zeta})] \right| \rightarrow 0 \text{ for all } \zeta \in \mathbb{R}. \tag{8}$$

Consider a change-of-variables by defining  $\mathbb{X}_c$  such that

$$\mathbb{X}_c(-u, v) = \mathbb{H}_n(ut_n, (v-u)t_n) \text{ for } (ut_n, (v-u)t_n) \in B_n. \tag{9}$$

The domain of  $\mathbb{X}_c$  is thus  $D_c \equiv \{(-u, v) \in (-t_n^{-1}\mathcal{X}) \times t_n^{-1}\mathcal{X} : v-u \geq \mathbf{1}\}$ . Note that  $\mathbb{X}_c(-u, v)$  is a normalized sum over observations for which  $x_i$  lies in the rectangle  $\{x | ut_n < x < vt_n\}$ . The change of variable and unusual notation are designed so that, for  $a, b \geq 0$ , the rectangle associated with  $\mathbb{X}_c(-u+a, v+b)$  contains the rectangle associated with  $\mathbb{X}_c(-u, v)$ . This helps with the verification of some of the conditions in Chan and Lai (2006) involving positive increments of the process.

Let  $\psi(z) = \frac{1}{z\sqrt{2\pi}}e^{-z^2/2}$  and  $\Delta_c = (2c^2)^{-1}$ . Consider a restriction of  $D_c$  to  $D_L (= D_{c,L}) \equiv \{(-u, v) \in D_c : v-u \leq L\mathbf{1}\}$  for some  $L > 1$ . Let

$$D_w^* = \{(-u, v) \in -I(w, |\log t_n|) \times I(w, |\log t_n|) : \mathbf{1} \leq (v-u) \leq L\mathbf{1}\}. \tag{10}$$

We will show that regularity conditions (C) and (A1)–(A5) in Corollary 2.7 of Chan and Lai (2006) are satisfied uniformly on the domains  $D_L$  and over  $P \in \mathcal{P}$  and hence

$$q_{w,P} \equiv P\left\{ \sup_{(-u,v) \in D_w^*} \mathbb{X}_c(-u, v) \geq c \right\} \sim \psi(c)\Delta_c^{-2d} \int_{D_w^*} H(-u, v) d(-u, v) \tag{11}$$

uniformly over  $I(w, |\log t_n|) \subseteq t_n^{-1}\mathcal{X}$  and  $P \in \mathcal{P}$ , where  $H$  is defined in that paper and, as shown below, takes the form

$$H(-u, v) = 4^{-2d} \text{vol}(v-u)^{-2} \tag{12}$$

in our case. Conditions (C) and (A1)–(A2) of Chan and Lai (2006) are verified in Section B.2, and conditions (A3)–(A5) are verified in Section B.5.

We partition  $t_n^{-1}\mathcal{X}$  into cubes of length  $|\log t_n|$  and apply (11) on each cube to show (8). More specifically, define  $Q_n = \{w \in (|\log t_n|\mathbb{Z})^d : I(w, |\log t_n|) \subseteq t_n^{-1}\mathcal{X}\}$ . Since  $\mathcal{X}$  is Jordan measurable and  $t_n|\log t_n| \rightarrow 0$ ,

$$\#Q_n \sim \text{vol}(\mathcal{X}) / (t_n|\log t_n|)^d, \quad (13)$$

and it follows from (7), (11) and (12) that

$$\sum_{w \in Q_n} q_{w,P} \rightarrow \lambda \equiv (1 - L^{-1})^d e^{-\zeta} \quad (14)$$

uniformly over  $P \in \mathcal{P}$ , noting that  $\lambda$  is the limit of  $\psi(c)\Delta_c^{-2d}(\#Q_n)|\log t_n|^d \int_{[0,L]^d} \text{vol}(t)^{-2} dt$ . Since  $\mathbb{X}_c$  is independent over  $D_{w_1}^*$  and  $D_{w_2}^*$  for  $w_1, w_2 \in Q_n$ ,  $w_1 \neq w_2$ , it follows from the Poisson limit of the Binomial distribution that

$$P\left\{ \sup_{w \in Q_n} \sup_{(-u,v) \in D_w^*} \mathbb{X}_c(-u, v) \geq c \right\} \rightarrow 1 - e^{-\lambda}$$

uniformly over  $P \in \mathcal{P}$ . Hence to show (8), it suffices for us to prove the following:

**Lemma B.4.** (a) *For all  $\epsilon > 0$ , there exists  $L$  large enough such that*

$$p_1 \equiv \sup_{P \in \mathcal{P}} P\left\{ \sup_{(-u,v) \in D_c \setminus D_{c,L}} \mathbb{X}_c(-u, v) \geq c \right\} \leq \epsilon \text{ for all large } c.$$

(b)  $p_2 \equiv \sup_{P \in \mathcal{P}} \sum_{w_1, w_2 \in Q_n, w_1 \neq w_2} P\left\{ \sup_{u \in I(w_1, |\log t_n|), v \in I(w_2, |\log t_n|), 1 \leq v - u \leq L} \mathbb{X}_c(-u, v) \geq c \right\} \rightarrow 0$ .

(c)  $p_3 \equiv \sup_{P \in \mathcal{P}} P\left\{ \sup_{(-u,v) \in D_L \setminus \cup_{w \in Q_n} I(w, |\log t_n|)} \mathbb{X}_c(-u, v) \geq c \right\} \rightarrow 0$ .

We prove this lemma in Section B.3. Sections B.2 and B.5 verify the conditions of Chan and Lai (2006) for the tail approximations used in the above argument. Section B.4 extends the results to multivariate  $Y_i$ .

## B.2 On (12) and the Verification of (C), (A1) and (A2)

Let  $\Phi$  be the c.d.f. of the standard normal.

**Lemma B.5.** (a) Let  $S_n = U_1 + \cdots + U_n$  and  $s_n^2 = \text{Var}(S_n)$ . Assume that  $U_1, \dots, U_n$  are independent mean 0 random variables and there exists  $\lambda > 0$ ,  $M_\lambda < \infty$  and  $\sigma_0^2 > 0$  such that

$$E(e^{\lambda|U_k|}) \leq M_\lambda, \quad \text{Var}(U_k) \geq \sigma_0^2, \quad 1 \leq k \leq n.$$

Let  $1 \leq x_n = o(n^{1/6})$ . Then there exists a constant  $C > 0$  dependent only on  $\lambda$ ,  $M_\lambda$ ,  $\sigma_0$  and  $\{x_n\}_{n \geq 1}$  such that

$$\left| \frac{P(S_n > xs_n)}{1 - \Phi(x)} - 1 \right| \leq \frac{Cx^3}{\sqrt{n}} \text{ for all } 1 \leq x \leq x_n, n \geq 1.$$

(b) [(A1) of Chan and Lai (2006)]  $P\{\mathbb{H}_n(s, t) \geq c - y/c\} \sim \psi(c - y/c)[\sim 1 - \Phi(c - y/c)]$  uniformly over  $P \in \mathcal{P}$  and positive, bounded values of  $y$  and  $(s, t) \in B_n$ .

*Proof.* The special case of i.i.d.  $U_k$  in (a) reduces to Theorem 1 in Chapter 16.6 of Feller (1971). Theorem 3 in Chapter 16.7 of Feller (1971) extends Theorem 1 to non-identically distributed random variables  $U_k$  such that  $E(|U_k|^3)/E(U_k^2)$  are uniformly bounded, with a  $O(\frac{x^3}{s_n})$  instead of a  $\frac{Cx^3}{\sqrt{n}}$  error bound. We follow step-by-step the proof of Feller's Theorem 3, using the additional condition  $\text{Var}(U_k) \geq \sigma_0^2$  to obtain the  $\frac{Cx^3}{\sqrt{n}}$  error bound in (a).

Under Assumption B.1,  $\mathbb{H}_n(s, t) = \frac{S^*}{\sigma_n(s, t)}$ , where  $S^*$  is a sum of independent mean 0 random variables satisfying (i) and (ii) with the bounds uniform over  $P \in \mathcal{P}$  and  $\text{Var}(S^*) = \sigma_n^2(s, t)$ . Hence by (a),

$$\frac{P\{\mathbb{H}_n(s, t) \geq c - y/c\}}{1 - \Phi(c - y/c)} = 1 + O\left(\frac{(c - y/c)^3}{\sigma_n(s, t)}\right) \text{ as } c - y/c \rightarrow \infty \quad (15)$$

uniformly over  $P \in \mathcal{P}$  and  $(s, t) \in B_n$ . Since  $\text{vol}(I(s, t)) \geq t_n^d$  for  $(s, t) \in B_n$ , by Assumption B.2(b),

$$\liminf_{n \rightarrow \infty} [\inf_{(s, t) \in B_n} \#J_n(s, t)] / (nt_n^d) \geq \inf_{x \in \mathcal{X}} f(x) > 0.$$

Hence by Assumption B.1(b) and B.2(a),  $\sigma_n^{-2}(s, t) = O((nt_n^d)^{-1}) = o(|\log t_n|^{-4})$  uniformly over  $(s, t) \in B_n$  and  $P \in \mathcal{P}$ . Since  $c = O(|\log t_n|^{1/2})$ , (b) follows from (15).  $\square$

Let  $\rho_c(-u, v, -u_1, v_1) = \text{Cov}(\mathbb{X}_c(-u, v), \mathbb{X}_c(-u_1, v_1))$  (we suppress the dependence of  $\rho_c$  on  $P$  in the notation) and let  $\{W_{-u, v}(q, r) : (q, r) \in [0, \infty)^{2d}\}$  be a continuous Gaussian

random field satisfying

$$W_{-u,v}(0) = 0, \quad E[W_{-u,v}(q, r)] = - \sum_{j=1}^d \frac{q_j + r_j}{4(v_j - u_j)}, \quad (16)$$

$$\text{Cov}(W_{-u,v}(q, r), W_{-u,v}(\alpha, \beta)) = \sum_{j=1}^d \frac{\min(q_j, \alpha_j) + \min(r_j, \beta_j)}{2(v_j - u_j)}.$$

**Lemma B.6.** (a) [(C) of Chan and Lai (2006)]  $1 - \rho_c(-u, v, -u + \delta_u, v + \delta_v) \sim \sum_{j=1}^d \frac{\delta_{u,j} + \delta_{v,j}}{2(v_j - u_j)}$  uniformly over  $(-u, v) \in D_L$  and  $P \in \mathcal{P}$  and compact sets of  $(\delta_u, \delta_v)/\Delta_c > 0$ .

(b) [(A2) of Chan and Lai (2006)] For any  $a > 0$  and positive integer  $m$ , as  $c \rightarrow \infty$ ,

$$\begin{aligned} & \{c[\mathbb{X}_c(-u + ak_u \Delta_c, v + ak_v \Delta_c) - \mathbb{X}_c(-u, v)] : 0 \leq (k_u, k_v) < m\mathbf{1}\} | \mathbb{X}_c(-u, v) = c - y/c \\ & \xrightarrow{d} \{W_{-u,v}(ak_u, ak_v) : 0 \leq (k_u, k_v) < m\mathbf{1}\}, \end{aligned}$$

uniformly over  $(-u, v) \in D_L$  and  $P \in \mathcal{P}$  and positive bounded values of  $y$ .

(c)  $H(-u, v) \equiv \lim_{K \rightarrow \infty} \int_0^\infty e^y P\{\sup_{\mathbf{0} \leq (q,r) \leq K\mathbf{1}} W_{-u,v}(q, r) \geq y\} dy$  has the closed-form given in (12).

*Proof.* Let  $z_0 = (s, t)$ , where  $s = ut_n$ ,  $t = (v - u)t_n$  and  $z_\delta = (s - \delta_u t_n, t + (\delta_v + \delta_u)t_n)$  for some  $\delta_u, \delta_v \geq 0$ . Then

$$\begin{aligned} \rho_c(-u, v, -u + \delta_u, v + \delta_v) &= \sigma_n(z_0)/\sigma_n(z_\delta) \\ &= \left(1 + \frac{\sigma_n^2(z_\delta) - \sigma_n^2(z_0)}{\sigma_n^2(z_0)}\right)^{-1/2} = 1 - \frac{\sigma_n^2(z_\delta) - \sigma_n^2(z_0)}{2\sigma_n^2(z_0)} + O\left(\frac{\sigma_n^2(z_\delta) - \sigma_n^2(z_0)}{\sigma_n^2(z_0)}\right)^2. \end{aligned} \quad (17)$$

Since  $\Delta_c \sim (4|\log t_n|)^{-1}$ , by Assumption B.2,

$$\sigma_n^2(z_0) \sim n\sigma^2(s)f(s)\text{vol}(t), \quad [\sigma_n^2(z_\delta) - \sigma_n^2(z_0)] \sim n\sigma^2(s)f(s)\text{vol}(t) \sum_{j=1}^d \frac{\delta_{u,j} + \delta_{v,j}}{v_j - u_j}, \quad (18)$$

and (a) follows from substituting (18) into (17).

Let  $a > 0$  and let  $\delta_u = ak_u \Delta_c$ ,  $\delta_v = ak_v \Delta_c$ . Then

$$c[\mathbb{X}_c(-u + \delta_u, v + \delta_v) - \mathbb{X}_c(-u, v)] = c\left\{\frac{\sum_{i \in J_n(z_\delta) \setminus J_n(z_0)} Z_i}{\sigma_n(z_\delta)} + \mathbb{X}_c(-u, v) \left[\frac{\sigma_n(z_0)}{\sigma_n(z_\delta)} - 1\right]\right\}. \quad (19)$$

We note here that as  $\delta_u, \delta_v \geq 0$ , so  $J_n(z_\delta) \supseteq J_n(z_0)$ . By (16)–(18), conditioned on  $\mathbb{X}_c(-u, v) =$

$c - y/c$  and noting that  $c^2 \Delta_c = \frac{1}{2}$ ,

$$c\mathbb{X}_c(-u, v) \left[ \frac{\sigma_n(z_0)}{\sigma_n(z_\delta)} - 1 \right] \rightarrow \sum_{j=1}^d \frac{\delta_{u,j} + \delta_{v,j}}{4(v_j - u_j)} = E[W_{-u,v}(ak_u, ak_v)], \quad (20)$$

$$\text{Var}\left(\frac{c \sum_{i \in J_n(z_\delta) \setminus J_n(z_0)} Z_i}{\sigma_n(z_\delta)}\right) = \frac{c^2[\sigma_n^2(z_\delta) - \sigma_n^2(z_0)]}{\sigma_n^2(z_\delta)} \rightarrow \text{Var}(W_{-u,v}(ak_u, ak_v)). \quad (21)$$

Similarly, if  $z_{\tilde{\delta}} = (s - \tilde{\delta}_u t_n, t + (\tilde{\delta}_u + \tilde{\delta}_v) t_n)$ , where  $\delta_u = \tilde{k}_u \Delta_c$ ,  $\delta_v = \tilde{k}_v \Delta_c$  with  $\tilde{k}_u, \tilde{k}_v \geq 0$ , then

$$\begin{aligned} \text{Cov}\left(\frac{c \sum_{i \in J_n(z_\delta) \setminus J_n(z_0)} Z_i}{\sigma_n(z_\delta)}, \frac{c \sum_{i \in J_n(z_{\tilde{\delta}}) \setminus J_n(z_0)} Z_i}{\sigma_n(z_{\tilde{\delta}})}\right) &= \frac{c^2[\sigma_n^2(z_{\min(\delta, \tilde{\delta})}) - \sigma_n^2(z_0)]}{\sigma_n(z_\delta) \sigma_n(z_{\tilde{\delta}})} \\ &\rightarrow \text{Cov}(W_{-u,v}(ak_u, ak_v), W_{-u,v}(a\tilde{k}_u, a\tilde{k}_v)). \end{aligned} \quad (22)$$

Since  $\sum_{i \in J_n(z_\delta) \setminus J_n(z_0)} Z_i$  is independent of  $\mathbb{X}_c(-u, v)$  and is asymptotically normal by Assumptions B.1(b)–(c) and B.2(b), (b) follows from (20)–(22). Lastly, (c) is a direct consequence of Lemma 2.3 of Chan and Lai (2006).  $\square$

### B.3 Proof of Lemma B.4

To deal with technicalities associated with non-rectangular edges, we extend the domain of  $\mathbb{H}_n$  to  $\mathcal{C} \times [t_n, 1]^d$  for some  $\mathcal{C} = [-C, C]^d$  by embedding  $(x_1, Y_1), (x_2, Y_2), \dots$  as a subsequence of  $(\tilde{x}_1, \tilde{Y}_1), (\tilde{x}_2, \tilde{Y}_2), \dots$  with  $\tilde{x}_i \in [-(C+1), (C+1)]^d$ . Hence the domain of  $\mathbb{X}_c$  can be extended to  $\{(-u, v) \in t_n^{-1} \mathcal{C}^2 : \mathbf{1} \leq v - u \leq t_n^{-1} \mathbf{1}\}$  with (C) and (A1)–(A5) satisfied uniformly over  $\{(-u, v) \in t_n^{-1} \mathcal{C}^2 : \mathbf{1} \leq v - u \leq L\mathbf{1}\}$  for any fixed  $L > 1$ .

*Proof of Lemma B.4(c).* Let  $\tilde{Q}_n = \{w \in (|\log t_n| \mathbb{Z})^d : I(w, |\log t_n|) \cap (t_n^{-1} \mathcal{X}) \neq \emptyset\}$ . Since  $\mathcal{X}$  is Jordan measurable,  $\#Q_n \sim \#\tilde{Q}_n$ . Hence by (11) and (14),

$$p_3 \leq \sup_{P \in \mathcal{P}} \sum_{w \in \tilde{Q}_n \setminus Q_n} q_{w,P} = o\left(\sup_{P \in \mathcal{P}} \sum_{w \in Q_n} q_{w,P}\right) = o(1).$$

$\square$

*Proof of Lemma B.4(b).* For  $n$  large enough such that  $|\log t_n| > L$ ,  $u \in I(w_1, |\log t_n|)$ ,  $v \in I(w_2, |\log t_n|)$ ,  $v - u \leq L\mathbf{1}$  can occur only when  $w_1, w_2$  are neighboring cubes. Note that

each cube has not more than  $3^d - 1$  neighbors. When  $w_1$  and  $w_2$  neighboring cubes, define

$$D_{w_1, w_2}^* = \{(-u, v) : u \in I(w_1, |\log t_n|), v \in I(w_2, |\log t_n|), \mathbf{1} \leq v - u \leq L\mathbf{1}\}.$$

Since  $\text{vol}(D_{w_1, w_2}^*) = o(|\log t_n|^{-d})$  and  $H(-u, v) \leq 1$ , by Corollary 2.7 of Chan and Lai (2006),

$$\begin{aligned} P\left\{ \sup_{(-u, v) \in D_{w_1, w_2}^*} \mathbb{X}_c(-u, v) \geq c \right\} &\sim \psi(c) \Delta_c^{-2d} \int_{D_{w_1, w_2}^*} H(-u, v) d(-u, v) \\ &= o(\psi(c) \Delta_c^{-2d} |\log t_n|^{-d}) \end{aligned}$$

uniformly over  $P \in \mathcal{P}$  and over neighboring  $w_1$  and  $w_2$ . Hence by (7) and (13),  $p_2 = o((\#Q_n) \psi(c) \Delta_c^{-2d} |\log t_n|^{-d}) = o(1)$ .  $\square$

*Proof of Lemma B.4(a).* For each  $\ell \in \mathbb{Z}^d$ ,  $\ell \neq 0$  with  $0 \leq \ell \leq [\log_L(2t_n^{-1}C)]\mathbf{1}$ , define

$$\mathbb{X}_{c, \ell}(-u, v) = \mathbb{H}_n(ut_n L^\ell, (v - u)t_n L^\ell) \text{ for } u, t \in (t_n L^\ell)^{-1}\mathcal{C} \text{ with } (v - u) \geq \mathbf{1}.$$

We use here the convention  $a\mathcal{C} = \prod_{j=1}^d [-a_j C, a_j C]$ . To avoid double counting, we restrict the domain of  $\mathbb{X}_{c, \ell}$  to

$$D_\ell \equiv \{(-u, v) \in (t_n L^\ell)^{-1}\mathcal{C}^2 : \mathbf{1} \leq v - u \leq L\mathbf{1}\}.$$

By Corollary 2.7 of Chan and Lai (2006),

$$P\left\{ \sup_{(-u, v) \in D_\ell} \mathbb{X}_c(-u, v) \geq c \right\} \sim \psi(c) \Delta_c^{-2d} \int_{D_\ell} H(-u, v) d(-u, v) \text{ uniformly over } \ell \text{ and } P \in \mathcal{P}, \quad (23)$$

with  $H(-u, v) = O(1)$  uniformly over  $\ell$  and  $D_\ell$ . By (7),  $\psi(c) \Delta_c^{-2d} \text{vol}(t_n^{-1}\mathcal{C}) = O(1)$  and so

$$\psi(c) \Delta_c^{-2d} \int_{D_\ell} H(-u, v) d(-u, v) = O(|L^\ell|^{-1}) \text{ uniformly over } \ell. \quad (24)$$

Hence by (23) and (24),

$$p_1 \leq \sup_{P \in \mathcal{P}} \sum_{0 \leq \ell \leq [\log_L(2t_n^{-1}C)]\mathbf{1}, \ell \neq 0} P\left\{ \sup_{(-u, v) \in D_\ell} \mathbb{X}_c(-u, v) \geq c \right\} = O\left( \sum_{0 \leq \ell \leq [\log_L(2t_n^{-1}C)]\mathbf{1}, \ell \neq 0} |L^\ell|^{-1} \right).$$

The sum above within  $O(\cdot)$  is bounded by  $(\sum_{k=0}^{\infty} L^{-k})^d - 1 = (1 - L^{-1})^{-d} - 1$  which can be made arbitrarily small by choosing  $L$  large enough.  $\square$

## B.4 Extension to Multivariate $Y$

Let  $E_{w,j} = \{\sup_{(-u,v) \in D_w^*} \mathbb{X}_{c,j} \geq c\}$ , where  $c = \frac{b_n + \min_{1 \leq j \leq d_Y} \zeta_j \sqrt{n}}{a_n}$ , for given  $\zeta_1, \dots, \zeta_{d_Y}$ , see (10). To extend the proof of Theorem B.1 to  $d_Y > 1$ , it suffices to prove the following:

**Lemma B.7.**  $p_4 \equiv \sup_{P \in \mathcal{P}} \sum_{w \in Q_n} \sum_{j_1 \neq j_2} P(E_{w,j_1} \cap E_{w,j_2}) \rightarrow 0$ .

*Proof.* Fix  $w$  and partition  $D_w^*$  into cubes of length  $\Delta_c$ . More specifically, define  $K_c = \{z \in (\Delta_c \mathbb{Z})^{2d} : I(z, \Delta_c \mathbf{1}) \cap D_w^* \neq \emptyset\}$ . Let  $G_{z,j} = \{\sup_{(-u,v) \in I(z, \Delta_c \mathbf{1})} \mathbb{X}_{c,j}(-u, v) \geq c\}$ . Then, uniformly over  $z, P \in \mathcal{P}$  and  $1 \leq j \leq d_Y$ ,

$$P(G_{z,j} \cap \{\mathbb{X}_{c,j}(z) \leq c - \theta/c\}) \sim \psi(c)H_\theta(z), \quad (25)$$

where  $H_\theta(z) = \int_\theta^\infty e^y P\{\sup_{0 \leq w \leq 1} W_z(w) > y\} dy$ .

This extends Theorem 2.4 of Chan and Lai (2006) to  $\theta \neq 0$ , using the same proof. Since  $H_0(z) < \infty$ , for any given  $\epsilon > 0$ , we can select  $\theta$  large enough such that  $H_\theta(z) \leq \epsilon$ . In addition, by (16), this selection can be made to be uniform over  $z \in K_c$  and  $1 \leq j \leq d_Y$ . Note that

$$P(E_{w,j_1} \cap E_{w,j_2}) \leq P(\mathbb{X}_{c,j_1}(z_1) > c - \theta/c, \mathbb{X}_{c,j_2}(z_2) > c - \theta/c \text{ for some } z_1, z_2 \in K_c) + \eta_{c,w},$$

where  $\eta_{c,w} = \sum_{z \in K_c} [P(G_{z,j_1} \cap \{\mathbb{X}_{c,j_1}(z) \leq c - \theta/c\}) + P(G_{z,j_2} \cap \{\mathbb{X}_{c,j_2}(z) \leq c - \theta/c\})]$ ,

and with  $\theta$  selected so that  $H_\theta(z) \leq \epsilon$ , it follows from (25) that

$$\eta_{c,w} = \epsilon O(\psi(c)(\#K_c)) = \epsilon O(\psi(c)\Delta_c^{-2d} |\log t_n|^d).$$

By (7),  $\psi(c) = O(t_n^d \Delta_c^{2d})$  and hence by (13),  $\sum_{w \in Q_n} \eta_{c,w} = \epsilon O(1)$ . It remains for us to show that for all  $\theta > 0$ ,

$$\sum_{z_1, z_2 \in K_n} P(\mathbb{X}_{c,j_1}(z_1) > c - \theta/c, \mathbb{X}_{c,j_2}(z_2) > c - \theta/c) = o(\psi(c)\Delta_c^{-2d} |\log t_n|^d). \quad (26)$$

Now by Assumption B.1(d),  $S(z_1, z_2) \equiv \mathbb{X}_{c,j_1}(z_1) \cap \mathbb{X}_{c,j_2}(z_2)$  has mean 0 and variance lying between  $2(1 - \rho)$  and  $2(1 + \rho)$ . Let  $\kappa = (\frac{2}{1+\rho})^{1/2} (> 1)$ . By Lemma B.5(a),

$$P\{S(z_1, z_2) \geq 2(c - \theta/c)\} \leq \left[1 + O\left(\frac{c^3}{\sqrt{nt_n^d}}\right)\right] [1 - \Phi(\kappa(c - \theta/c))] \sim \frac{1}{\kappa c (2\pi)^{1/2}} e^{-\kappa^2 c^2 / 2 + \kappa \theta} \quad (27)$$

uniformly over  $P \in \mathcal{P}$ . Since  $\#K_c = O(\Delta_c^{-2d} |\log t_n|^d)$ , it follows from (36) that

$$\begin{aligned} & \sum_{z_1, z_2 \in K_c} P(\mathbb{X}_{c,j_1}(z_1) > c - \theta/c, \mathbb{X}_{c,j_2}(z_2) > c - \theta/c) \\ & \leq \sum_{z_1, z_2 \in K_c} P\{S(z_1, z_2) > 2(c - \theta/c)\} = O(\psi(c)e^{-(\kappa^2-1)c^2/2} \Delta_c^{-4d} |\log t_n|^{2d}) \end{aligned}$$

uniformly over  $P \in \mathcal{P}$  and (26) holds because  $|\log t_n|^d = O(c^{2d})$  and  $c^{6d}e^{-(\kappa^2-1)c^2/2} = o(1)$ .  $\square$

## B.5 Verification of (A3)–(A5)

Conditions (C), (A1) and (A2) have been verified in Section B.2. The remaining regularity conditions that lead to (11) will be verified in Lemmas B.8 and B.9 below.

**Lemma B.8.** (a) [(A3) of Chan and Lai (2006)] *Let  $\gamma > 0$  and  $k_u, k_v \geq 0$ . There exists a positive function  $h$  such that  $\lim_{y \rightarrow \infty} h(y) = 0$  and*

$$P\{\mathbb{X}_c(-u + k_u \Delta_c, v + k_v \Delta_c) > c - \gamma/c, \mathbb{X}_c(-u, v) \leq c - y/c\} \leq h(y)\psi(c) \text{ for all large } c,$$

uniformly over  $(-u, v) \in D_L$  and  $P \in \mathcal{P}$ .

(b) [(A5) of Chan and Lai (2006)] *There exists a nonincreasing positive function  $r$  on  $[0, \infty)$  such that  $r(\|k\|) = O(e^{-\|k\|^p})$  for some  $p > 0$  such that for any  $\gamma > 0$ ,*

$$P\{\mathbb{X}_c(-u, v) > c - \gamma/c, \mathbb{X}_c(-u + k_u \Delta_c, v + k_v \Delta_c) > c - \gamma/c\} \leq \psi(c - \gamma/c)r(\|k_u, k_v\|) \text{ for all large } c,$$

uniformly over  $P \in \mathcal{P}$ ,  $(-u, v), (-u + k_u \Delta_c, v + k_v \Delta_c) \in D_w^*$  and  $w \in Q_n$ .

*Proof.* Let  $\omega > 1$  to be specified later. By Lemma B.5(a), there exists  $\xi_c \rightarrow 0$  such that

$$P\{\mathbb{X}_c(-u, v) \geq c - y'/c\} = [1 + O(\xi_c^2)]e^{y'}\psi(c)$$

uniformly over  $\gamma \leq y' \leq \omega c$  and  $P \in \mathcal{P}$ . Let  $y_j = y + j\xi_c$ ,  $j = 0, 1, \dots$ . Let  $u_1 = u - k_u \Delta_c$  and  $v_1 = v + k_v \Delta_c$ . Since  $e^{\xi_c} = 1 + \xi_c + O(\xi_c^2)$ ,

$$\begin{aligned} & P\{\mathbb{X}_c(-u, v) > c - y_{j+1}/c\} - P\{\mathbb{X}_c(-u, v) > c - y_j/c\} \\ & = [1 + O(\xi_c^2)]e^{y_j + \xi_c}\psi(c) - [1 + O(\xi_c^2)]e^{y_j}\psi(c) \sim \xi_c e^{y_j}\psi(c) \end{aligned} \tag{28}$$



uniformly over  $\gamma \leq y_j \leq \omega c$  and  $P \in \mathcal{P}$ . Since  $P\{\mathbb{X}_c(-u_1, v_1) > a | \mathbb{X}_c(-u, v) = b\}$  increases with  $b$  for any fixed  $a$ , it follows from (28) that

$$\begin{aligned}
& P\{\mathbb{X}_c(-u_1, v_1) > c - y/c, c - \omega \leq \mathbb{X}_c(-u, v) < c - y/c\} \\
& \leq \sum_{0 \leq j \leq (\omega c - y)/\xi_c} P\{\mathbb{X}_c(-u_1, v_1) > c - \gamma/c | \mathbb{X}_c(-u, v) = c - y_j/c\} \\
& \quad \times [P\{\mathbb{X}_c(-u, v) > c - y_{j+1}/c\} - P\{\mathbb{X}_c(-u, v) > c - y_j/c\}] \\
& \sim \psi(c)\xi_c \sum_{0 \leq y_j \leq \omega c} e^{y_j} P\{\mathbb{X}_c(-u_1, v_1) > c - \gamma/c | \mathbb{X}_c(-u, v) = c - y_j/c\}.
\end{aligned} \tag{29}$$

Let  $z_\delta = (u_1 t_n, (v_1 - u_1) t_n)$ ,  $k = (k_u, k_v)$  and let

$$g_{-u,v}(k) = \sum_{j=1}^d \frac{k_{u,j} + k_{v,j}}{4(v_j - u_j)} \{= E[W_{-u,v}(k_u, k_v)] = \text{Var}(W_{-u,v}(k_u, k_v))/2\}, \tag{30}$$

(see (16)). Then by (18)–(21) with  $a = 1$ ,

$$\begin{aligned}
& P\{c[\mathbb{X}_c(-u_1, v_1) - \mathbb{X}_c(-u, v)] \geq y_j - \gamma | \mathbb{X}_c(-u, v) = c - y_j/c\} \\
& = P\left\{ \frac{\sum_{i \in J_n(z_\delta) \setminus J_n(z_0)} Z_i}{\sqrt{\sigma_n^2(z_\delta) - \sigma_n^2(z_0)}} \geq \frac{y_j - \gamma - c(c - y_j/c)[\frac{\sigma_n(z_0)}{\sigma_n(z_\delta)} - 1]\sigma_n(z_\delta)}{c\sqrt{\sigma_n^2(z_\delta) - \sigma_n^2(z_0)}} \right\} \\
& = P\left\{ \frac{\sum_{i \in J_n(z_\delta) \setminus J_n(z_0)} Z_i}{\sqrt{\sigma_n^2(z_\delta) - \sigma_n^2(z_0)}} \geq \frac{y_j - \gamma + g_{-u,v}(k) + o(1)}{\sqrt{2g_{-u,v}(k) + o(1)}} \right\} \sim \psi\left(\frac{y_j - \gamma + g_{-u,v}(k)}{\sqrt{2g_{-u,v}(k)}}\right),
\end{aligned} \tag{31}$$

with  $o(1)$  uniform over  $y \leq y_j \leq \omega c$  and  $(-u, v) \in D_L$  and  $P \in \mathcal{P}$ , noting that as  $y' = O(c)$ , the relative error of the normal tail approximation in (31) is

$$O\left(\frac{c^3}{\sqrt{\sigma_n^2(z_\delta) - \sigma_n^2(z_0)}}\right) = O\left(\frac{c^4}{\sqrt{nt_n^d}}\right) \rightarrow 0$$

(see Assumption B.2(a) and Lemma B.5(a)). By (29) and (31),

$$\begin{aligned}
& P\{\mathbb{X}_c(-u_1, v_1) > c - \gamma/c, c - \omega < \mathbb{X}_c(-u, v) \leq c - y/c\} \\
& \sim \psi(c) \int_y^{\omega c} e^{y'} \psi\left(\frac{y' - \gamma + g_{-u,v}(k)}{\sqrt{2g_{-u,v}(k)}}\right) dy'.
\end{aligned} \tag{32}$$

To complete the proof of (a), it suffices to show that

$$(II) \equiv P\{\mathbb{X}_c(-u_1, v_1) > c - \gamma/c, \mathbb{X}_c(-u, v) \leq c - \omega\} = o(\psi(c)) \tag{33}$$

for all  $\omega$  large. By (6) and (9),

$$\sum_{i \in J_n(z_\delta) \setminus J_n(z_0)} Z_i = \sigma_n(z_\delta) \mathbb{X}_c(-u_1, v_1) - \sigma_n(z_0) \mathbb{X}_c(-u, v). \quad (34)$$

Since  $\sigma_n(z_\delta) \geq \sigma_n(z_0)$ , by (21), (30) and Lemma B.5(a) with  $a = 1$ ,

$$(II) \leq P \left\{ \frac{\sum_{i \in J_n(z_\delta) \setminus J_n(z_0)} Z_i}{\sqrt{\sigma_n^2(z_\delta) - \sigma_n^2(z_0)}} \geq \frac{\sigma_n(z_0)(\omega c - \gamma)}{c \sqrt{\sigma_n^2(z_\delta) - \sigma_n^2(z_0)}} \right\} = \left[ 1 + O \left( \frac{x_n^3}{\sqrt{nt_n^d}} \right) \right] \psi(x_n),$$

where  $x_n = \frac{\omega c - \gamma}{c \sqrt{4\Delta_c [g_{-u,v}(k) + o(1)]}} = \frac{\omega c - \gamma}{\sqrt{2[g_{-u,v}(k) + o(1)]}}$ ,

and indeed (33) holds when  $\omega > \sqrt{2g_{-u,v}(k)}$ .

To prove (b), we apply Lemma B.5(a) to the right-hand side of the inequality

$$P\{\mathbb{X}_c(-u, v) > c - \gamma/c, \mathbb{X}_c(-u_1, v_1) > c - \gamma/c\} \leq P\{\mathbb{X}_c(-u, v) + \mathbb{X}_c(-u_1, v_1) > 2(c - \gamma/c)\} \equiv (III).$$

As in the proof of Lemma B.4(b), the relative error of the normal approximation goes to 0 due to Assumption B.2(a), that is,

$$(III) \sim \psi \left( \frac{2(c - \gamma/c)}{\sqrt{2 + 2\rho_c(-u, v, -u_1, v_1)}} \right) \text{ as } c \rightarrow \infty. \quad (35)$$

Note that in the statement of (b), the restriction  $k_u, k_v \geq 0$  is removed and we have in place of (17),

$$\rho_c(-u, v, -u_1, v_1) = \frac{\sigma_n^2(z^*)}{\sigma_n(z_0)\sigma_n(z_\delta)},$$

where  $z^* = (-(u \vee u_1), (v \wedge v_1))$ . Since  $J_n(z^*) = J_n(z_0) \cap J_n(z_\delta)$ , so by expanding  $\sigma_n(z^*)/\sigma_n(z_0)$  and  $\sigma_n(z^*)/\sigma_n(z_\delta)$  as in (17), it follows from (18) and (30) with  $\delta_u = k_u \Delta_c$ ,  $\delta_v = k_v \Delta_c$  that

$$\begin{aligned} \rho_c(-u, v, -u_1, v_1) &= 1 - (1 + o(1)) \left\{ \frac{\sigma_n^2(z_\delta) - \sigma_n^2(z^*)}{2\sigma_n^2(z^*)} + \frac{\sigma_n^2(z_0) - \sigma_n^2(z^*)}{2\sigma_n^2(z^*)} \right\} \\ &= 1 - (1 + o(1)) \left\{ \sum_{j=1}^d \frac{(\delta_{u,j})^+ + (\delta_{v,j})^+}{v_j - u_j} + \sum_{j=1}^d \frac{(\delta_{u,j})^- + (\delta_{v,j})^-}{v_j - u_j} \right\} \\ &= 1 - (4 + o(1)) \Delta_c g_{-u,v}(|k|), \end{aligned}$$

from which it follows that

$$\frac{2(c - \gamma/c)}{\sqrt{2 + 2\rho_c(-u, v, -u_1, v_1)}} = (c - \gamma/c)[1 - (2 + o(1))\Delta_c g_{-u, v}(|k|)]^{-1/2} \geq c + [g_{-u, v}(|k|)/2 - \gamma + o(1)]/c,$$

and (b) with  $r(\tau) = \exp[-\min_{\|k\|=\tau} g_{-u, v}(|k|)/4]$  and  $0 < p < 1$ , follows from (35).  $\square$

**Lemma B.9.** (a) [Theorem 1 of Wichura (1969)] *Let  $\mathcal{A}$  be a finite subset of  $\mathbb{R}^d$  and let  $U_i$ ,  $i \in \mathcal{A}$  be independent mean 0 random variables with variance  $\sigma_i^2$ . Let  $S_k = \sum_{i \leq k} U_i$  and set  $s_{\mathcal{A}}^2 = \sum_{i \in \mathcal{A}} \sigma_i^2$ ,  $S_{\mathcal{A}} = \sum_{i \in \mathcal{A}} U_i$ . Then for any  $x > 2^d s_{\mathcal{A}}$ ,*

$$P(\max_{k \in \mathbb{R}^d} |S_k| > x) \leq [1 - (2^d s_{\mathcal{A}}/x)^2]^{-d} P(|S_{\mathcal{A}}| > 2^{-d} x). \quad (36)$$

(b) [(A4) of Chan and Lai (2006)] *There exists nonincreasing functions  $N_a$  on  $\mathbb{R}^+$  and positive constants  $\gamma_a \rightarrow 0$  such that  $N_a(\gamma_a) + \int_1^\infty \tau^s N_a(\gamma_a + \tau) d\tau = o(a^d)$  as  $a \rightarrow 0$  for all  $s > 0$ , and for each  $a > 0$ ,*

$$P\left\{ \sup_{0 \leq (k_u, k_v) \leq a\mathbf{1}} \mathbb{X}_c(-u + k_u \Delta_c, v + k_v \Delta_c) > c, \mathbb{X}_c(-u, v) \leq c - \gamma/c \right\} \leq N_a(\gamma) \psi(c), \quad (37)$$

uniformly over  $(-u, v) \in D_L$  and  $P \in \mathcal{P}$  for all  $\gamma_a \leq \gamma \leq c$  with  $c$  large.

*Proof.* Though Wichura (1969) considers a set  $\mathcal{A}$  with points lying on a  $d$ -dimensional grid, we can always extend  $\mathcal{A}$  to a  $d$ -dimensional grid  $\mathcal{B}$  by letting  $U_i \equiv 0$  for  $i \in \mathcal{B} \setminus \mathcal{A}$ . Note that the right-hand side of (36) is unchanged by such an extension. Let  $u_1 = u - k_u \Delta_c$ ,  $v_1 = v + k_v \Delta_c$ ,  $k = (k_u, k_v)$  and  $z_\delta = (u_1 t_n, (v_1 - u_1) t_n)$ . Let  $\omega > 1$  to be specified later. Since  $\sigma_n(z_\delta) \geq \sigma_n(z_0)$  when  $J_n(z_\delta) \supseteq J_n(z_0)$ , by the arguments in (29) and (31),

$$\begin{aligned} \text{(I)} &\equiv P\left\{ \sup_{0 \leq k \leq a\mathbf{1}} \mathbb{X}_c(-u_1, v_1) > c, c - \omega \leq \mathbb{X}_c(-u, v) \leq c - \gamma/c \right\} \\ &\sim \psi(c) \int_\gamma^{\omega c} e^y P\left\{ \sup_{0 \leq k \leq a\mathbf{1}} \sum_{i \in J_n(z_\delta) \setminus J_n(z_0)} Z_i \geq \sigma_n(z_0) y / c \right\} dy. \end{aligned} \quad (38)$$

Let  $\mathcal{B}$  be the set of all  $d$ -dimensional vectors with coordinates taking values  $-1, 0$  or  $1$  but not all zeros. Hence  $\#\mathcal{B} = 3^d - 1$ . Consider the partitioning of  $\mathcal{A} \equiv J_n((u - a\Delta_c)t_n, (v - u + 2a\Delta_c)t_n) \setminus J_n(z_0)$  as  $\mathcal{A} = \bigcup_{b \in \mathcal{B}} \mathcal{A}_b$ , with

$$\begin{aligned} \mathcal{A}_b &= \{i : (u_j - a\Delta_c)t_n \leq x_{i,j} < u_j t_n \text{ if } b_j = -1, \\ &\quad u_j t_n \leq x_{i,j} \leq v_j t_n \text{ if } b_j = 0, \end{aligned}$$

$$v_j t_n \leq x_{i,j} \leq (v_j + a\Delta_c)t_n \text{ if } b_j = 1\}.$$

Then

$$\sup_{0 \leq k \leq a\mathbf{1}} \sum_{i \in J_n(z_\delta) \setminus J_n(z_0)} Z_i \leq \sum_{b \in \mathcal{B}} \max_{k \in \mathbb{R}^d} S_{k,b}, \text{ where } S_{k,b} = \sum_{i \in \mathcal{A}_b, i \leq k} Z_i.$$

Let  $x = \frac{\sigma_n(z_0)y}{3^d c}$ . By (18), and since  $v_j - u_j \geq 1$ ,

$$\begin{aligned} \frac{x}{s_{\mathcal{A}}} &\sim \frac{y \text{vol}(v-u)}{3^d c (\text{vol}(v-u+2a\Delta_c \mathbf{1}) - \text{vol}(v-u))^{1/2}} = \frac{y}{3^d c} \left[ \prod_{j=1}^d \left(1 + \frac{2a\Delta_c}{v_j - u_j}\right) - 1 \right]^{-1/2} \quad (39) \\ &\geq (1+o(1)) \frac{y}{3^d c} \left(\frac{da}{c^2}\right)^{-1/2} \sim \frac{y}{3^d \sqrt{da}}. \end{aligned}$$

Hence for all large  $c$ ,

$$\frac{x}{s_{\mathcal{A}}} \geq 2^{d+1/2} \text{ when } y \geq \gamma \text{ for } a \leq \left(\frac{\gamma}{6^{d+1/2}}\right)^2 \frac{1}{d}.$$

By (36) and (38), and since  $s_{\mathcal{A}_b} \leq s_{\mathcal{A}}$ ,

$$(I) \leq (2^d + o(1)) \psi(c) \sum_{b \in \mathcal{B}} \int_{\gamma}^{\omega c} e^y P\left\{|S_{\mathcal{A}_b}| \geq \frac{\sigma_n(z_0)y}{6^d c}\right\} dy. \quad (40)$$

Apply Lemma B.5(a) and note that the sum in (40) is dominated by the  $2d$  values of  $b$  having a single non-zero entry. Then by (39), (37) holds (for large  $c$  and small  $a$ ) when

$$N_a(\gamma) = d2^{d+2} \int_{\gamma}^{\infty} e^y \psi\left(\frac{y}{6^{d+1/2} \sqrt{da}}\right) dy.$$

We check that when  $\gamma_a = a^{1/3}$ , then  $N_a(\gamma_a) + \int_1^{\infty} \tau^s N_a(\gamma_a + \tau) = o(a^p)$  as  $a \rightarrow 0$  for all  $s > 0$  and  $p > 0$ , and that

$$P\left\{\sup_{0 \leq k \leq a\mathbf{1}} \mathbb{X}_c(-u_1, v_1) > c, \mathbb{X}_c(-u, v) \leq c - \omega\right\} \leq P\left\{\sup_{0 \leq k \leq a\mathbf{1}} \sum_{i \in J_n(z_\delta) \setminus J_n(z_0)} Z_i \geq \omega \sigma_n(z_0)\right\} = o(\psi(c))$$

for all  $\omega$  large, by a similar partitioning argument and applications of Lemmas B.5(a) and B.9(a).  $\square$

## C Proof of Theorem 4.1

We have

$$\begin{aligned}\bar{m}_j(\theta_n, x) &= \bar{m}_j(\theta_0, x_0) + \bar{m}_j(\theta_n, x) - \bar{m}_j(\theta_0, x) + \bar{m}_j(\theta_0, x) - \bar{m}_j(\theta_0, x_0) \\ &= [\bar{m}_{\theta,j}(\theta_n^*, x)a]r_n + \bar{m}_j(\theta_0, x) - \bar{m}_j(\theta_0, x_0) \leq [\bar{m}_{\theta,j}(\theta_n^*, x)a]r_n + C \left( \frac{x - x_0}{\|x - x_0\|} \right) \|x - x_0\|^\gamma\end{aligned}$$

for  $x$  in some neighborhood of  $x_0$ . Thus, letting  $h$  be some small scalar going to zero with  $n$ , for  $sh$  and  $(s+t)h$  small enough, we have

$$\begin{aligned}& Em_j(W_i, \theta_n)I(sh + x_0 < X_i < (s+t)h + x_0) \\ & \leq \int_{x \in \mathcal{X}, sh+x_0 < x < (s+t)h+x_0} \left\{ [\bar{m}_{\theta,j}(\theta_n^*, x)a]r_n + C \left( \frac{x - x_0}{\|x - x_0\|} \right) \|x - x_0\|^\gamma \right\} f(x) dx \\ & = \int_{x_0+uh \in \mathcal{X}, s < u < s+t} \left\{ [\bar{m}_{\theta,j}(\theta_n^*, x_0 + uh)a]r_n/h^\gamma + C \left( \frac{u}{\|u\|} \right) \|u\|^\gamma \right\} f(x_0 + uh)h^{d_x+\gamma} du\end{aligned}$$

where the last equality uses the change of variables  $u = (x - x_0)/h$ . We also have

$$\begin{aligned}& \sigma_j^2(sh + x_0, (s+t)h + x_0, \theta_n) \\ & = Em_j(W_i, \theta_n)^2 I(sh + x_0 < X_i < (s+t)h + x_0) - [Em_j(W_i, \theta_n)I(sh + x_0 < X_i < (s+t)h + x_0)]^2 \\ & \leq \int_{x \in \mathcal{X}, sh+x_0 < x < (s+t)h+x_0} \mu_{2,j}^2(x)f(x) dx = \int_{x_0+uh \in \mathcal{X}, s < u < s+t} \mu_{2,j}^2(x_0 + uh)f(x_0 + uh)h^{d_x} dx.\end{aligned}$$

using the same change of variables. Under these assumptions,  $\mu_{2,j}^2(x_0 + uh)$  converges to  $\Sigma_{jj}(x_0)$  and  $f(x_0 + uh) \rightarrow f(x_0)$  uniformly over bounded  $u$  as  $h$  approaches 0.

Thus, for any  $\varepsilon > 0$ , we will have, for small enough  $h$  and bounded  $s$  and  $t$ ,  $Em_j(W_i, \theta_n)I(sh + x_0 < X_i < (s+t)h + x_0)/\sigma_j^2(sh + x_0, th, \theta_n)$  is, for any  $s, t$  such that the expression is negative, bounded from above by

$$h^{d_x/2+\gamma} \frac{[f(x_0)^{1/2} - \varepsilon] \int_{x_0+uh \in \mathcal{X}, s < u < s+t} \left\{ [\bar{m}_{\theta,j}(\theta_0, x_0)a + \varepsilon]r_n/h^\gamma + C \left( \frac{u}{\|u\|} \right) \|u\|^\gamma \right\} du}{[\Sigma_{jj}^{1/2}(x) + \varepsilon] \text{vol}\{u|x_0 + uh \in \mathcal{X}, s < u < s+t\}^{1/2}}$$

Setting  $h = r_n^{1/\gamma}$ , this is equal to

$$r_n^{(d_x/2+\gamma)/\gamma} \lambda(s, t, (\mathcal{X} - x_0)/r_n^\gamma, \varepsilon)$$

for a function  $\lambda$  that does not depend on  $r_n$ . Note that the sequence of sets  $(\mathcal{X} - x_0)/r_n^\gamma$  satisfies  $(\mathcal{X} - x_0)/r_k^\gamma \subseteq (\mathcal{X} - x_0)/r_\ell^\gamma$  for  $r_\ell < r_k$  by convexity of  $\mathcal{X}$ , so, letting  $\mathcal{U} = \cup_{k=1}^\infty (\mathcal{X} - x_0)/r_k$ , we will have  $\text{vol}(\{s < u < s+t\} \cap (U \setminus (\mathcal{X} - x_0)/r_k)) \rightarrow 0$ . It follows that  $\lambda(s, t, (\mathcal{X} - x_0)/r_n, \varepsilon) \rightarrow \lambda(s, t, \mathcal{U}, \varepsilon)$ . Since this holds for all  $\varepsilon$  and  $\lambda$  is continuous in  $\varepsilon$ , we have, for any  $s, t$ ,

$$\begin{aligned}
& - \frac{E_n m_j(W_i, \theta_n) I(sr_n^{1/\gamma} + x_0 < X_i < (s+t)r_n^{1/\gamma} + x_0)}{\hat{\sigma}_j(sr_n^{1/\gamma} + x_0, tr_n^{1/\gamma}, \theta_n)} - \hat{q}_{1-\alpha} \\
& = - \frac{E m_j(W_i, \theta_n) I(sr_n^{1/\gamma} + x_0 < X_i < (s+t)r_n^{1/\gamma} + x_0)}{\sigma_j(sr_n^{1/\gamma} + x_0, tr_n^{1/\gamma}, \theta_n)} + \mathcal{O}_P(n^{-1/2}) - \hat{q}_{1-\alpha} \\
& \geq r_n^{(d_X/2+\gamma)/\gamma} [-\lambda(s, t, \mathcal{U}) + o(1) - \hat{q}_{1-\alpha}/r_n^{(d_X/2+\gamma)/\gamma}] + \mathcal{O}_P(n^{-1/2}) \\
& = r_n^{(d_X/2+\gamma)/\gamma} \left\{ -\lambda(s, t, \mathcal{U}) + o(1) - \frac{(2 \log t_n^{-d_X})^{1/2}}{n^{1/2} r_n^{(d_X/2+\gamma)/\gamma}} (1 + o_P(1)) + \mathcal{O}_P(n^{-1/2} r_n^{-(d_X/2+\gamma)/\gamma}) \right\}
\end{aligned}$$

Note that the  $o_P(1)$  term absorbs the  $\mathcal{O}_P(1)$  term, so that the above expression will be negative with probability approaching one for some  $s, t$  as long as

$$\limsup \frac{(2 \log t_n^{-d_X})^{1/2}}{n^{1/2} r_n^{(d_X/2+\gamma)/\gamma}} < \sup_{s,t} -\lambda(s, t, \mathcal{U}),$$

and this condition can be rearranged to

$$\liminf r_n \left( \frac{n}{2 \log t_n^{-d_X}} \right)^{\gamma/(d_X+2\gamma)} > -1/[\inf_{s,t} \lambda(s, t, \mathcal{U})]^{\gamma/(d_X+2\gamma)}.$$

## D Comparison to Intermediate Gaussian Approximations

In this section of the appendix, we compare our approach to the results that could be obtained using intermediate gaussian approximations. As shown in Section 4,  $t_n$  must be chosen at least as small as the optimal bandwidth in order for the test to have good power for a given data generating process. Theorem 2.1 allows  $t_n$  to be chosen equal to  $n^{-1/d_X}$  times a  $\log n$  term, which is small enough to adapt to any Holder class for the conditional mean. Using the best available results for gaussian approximations in Rio (1994) would give a rate of approximation of a  $\log n$  term times  $n^{-1/[2(d_X+1)]}$  for the random process  $(s, t) \mapsto \sqrt{n} E_n m(W_i, \theta) I(s < X_i < s+t)$ . The test statistic weights this by the inverse of its estimated

standard deviation which, at the minimum scale  $t_n$ , is of order  $t_n^{-d_X/2}$ . Thus, in order to use the gaussian approximation of Rio (1994), we would need  $t_n^{-d_X/2} \cdot n^{-1/[2(d_X+1)]}$  to go to zero more quickly than a  $\log n$  term, which would mean that  $t_n$  would have to decrease more slowly than a  $\log n$  term times  $n^{-\frac{1}{d_X(d_X+1)}}$ . For the test to achieve optimal power when the conditional mean has  $\gamma$  continuous derivatives (where noninteger values of  $\gamma$  correspond to Holder conditions),  $t_n$  must decrease at least as quickly as  $n^{-1/(d_X+2\gamma)}$ . Thus, using a gaussian approximation would only lead to optimal power when  $\frac{1}{d_X+2\gamma} \leq \frac{1}{d_X(d_X+1)}$ , which can be rewritten as  $d_X + 2\gamma \geq d_X^2 + d_X$  or  $\gamma \geq d_X^2/2$ . Another approach is to restrict the set  $(s, t)$  over which the supremum is taken to a finite set and place conditions on the rate at which this set increases with the sample size. While this approach does not apply directly to the statistic considered here, it is useful to compare our results to this approach as well. Using the results of Chatterjee (2005) along with this approach and a method of proof that avoids deriving an asymptotic distribution, Chetverikov (2012) provides a test that is adaptive in the range  $\gamma \in [d_X, 2]$ .

Since the use of positive kernels (in this case indicator functions) prevents multiscale statistics from being adaptive to  $\gamma > 2$  derivatives, this means that the approach based on the gaussian approximations in Rio (1994) would be adaptive to a range of  $[d_X^2/2, 2]$  for the smoothness parameter  $\gamma$ . Thus, while this approach would lead to useful (if not optimal) results for a one dimensional covariate, it would not be adaptive in two dimensions, and would be dominated by a kernel statistic with a single bandwidth in more than two dimensions. In contrast, our result allows adaptivity to all  $\gamma$  in  $(0, 2]$  regardless of the dimension of  $X_i$ , which is the best possible result.

## References

- ANDREWS, D. W., S. BERRY, AND P. JIA (2004): “Confidence regions for parameters in discrete games with multiple equilibria, with an application to discount chain store location,” .
- ANDREWS, D. W., AND P. GUGGENBERGER (2009): “Validity of Subsampling and ?plug-in Asymptotic? Inference for Parameters Defined by Moment Inequalities,” *Econometric Theory*, 25(03), 669–709.
- ANDREWS, D. W., AND X. SHI (2009): “Inference Based on Conditional Moment Inequalities,” *Unpublished Manuscript, Yale University, New Haven, CT*.

- ANDREWS, D. W. K., AND P. JIA (2008): “Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure,” *SSRN eLibrary*.
- ANDREWS, D. W. K., AND G. SOARES (2010): “Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection,” *Econometrica*, 78(1), 119–157.
- ARMSTRONG, T. (2011a): “Asymptotically Exact Inference in Conditional Moment Inequality Models,” *Unpublished Manuscript*.
- (2011b): “Weighted KS Statistics for Inference on Conditional Moment Inequalities,” *Unpublished Manuscript*.
- (2012): “On the Choice of Test Statistic for Conditional Moment Inequality Models,” *Unpublished Manuscript*.
- BERESTEANU, A., AND F. MOLINARI (2008): “Asymptotic Properties for a Class of Partially Identified Models,” *Econometrica*, 76(4), 763–814.
- BICKEL, P. J., AND M. ROSENBLATT (1973): “On some global measures of the deviations of density function estimates,” *The Annals of Statistics*, p. 10711095.
- BUGNI, F. A. (2010): “Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set,” *Econometrica*, 78(2), 735–753.
- CHAN, H. P., AND T. L. LAI (2006): “Maxima of asymptotically Gaussian random fields and moderate deviation approximations to boundary crossing probabilities of sums of random variables with multidimensional indices,” *The Annals of Probability*, 34(1), 80–121.
- CHATTERJEE, S. (2005): “A simple invariance theorem,” *arXiv:math/0508213*.
- CHERNOZHUKOV, V., H. HONG, AND E. TAMER (2007): “Estimation and Confidence Regions for Parameter Sets in Econometric Models,” *Econometrica*, 75(5), 1243–1284.
- CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2009): “Intersection bounds: estimation and inference,” *Arxiv preprint arXiv:0907.3503*.
- CHEVRIKOV, D. (2012): “Adaptive Test of Conditional Moment Inequalities,” .



- CILIBERTO, F., AND E. TAMER (2009): “Market structure and multiple equilibria in airline markets,” *Econometrica*, 77(6), 1791–1828.
- DUMBGEN, L., AND V. G. SPOKOINY (2001): “Multiscale Testing of Qualitative Hypotheses,” *The Annals of Statistics*, 29(1), 124–152.
- FELLER, W. (1971): *An Introduction to Probability Theory and Its Applications, Vol. 2*. Wiley, volume 2 edn.
- HOROWITZ, J. L., AND V. G. SPOKOINY (2001): “An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model Against a Nonparametric Alternative,” *Econometrica*, 69(3), 599–631.
- IMBENS, G. W., AND C. F. MANSKI (2004): “Confidence Intervals for Partially Identified Parameters,” *Econometrica*, 72(6), 1845–1857.
- KHAN, S., AND E. TAMER (2009): “Inference on endogenously censored regression models using conditional moment inequalities,” *Journal of Econometrics*, 152(2), 104–119.
- KIM, K. I. (2008): “Set estimation and inference with models characterized by conditional moment inequalities,” .
- LEE, S., O. LINTON, AND Y.-J. WHANG (2009): “Testing for Stochastic Monotonicity,” *Econometrica*, 77(2), 585–602.
- LEE, S., K. SONG, AND Y. WHANG (2011): “Testing functional inequalities,” .
- MANSKI, C. F. (1990): “Nonparametric Bounds on Treatment Effects,” *The American Economic Review*, 80(2), 319–323.
- MANSKI, C. F., AND E. TAMER (2002): “Inference on Regressions with Interval Data on a Regressor or Outcome,” *Econometrica*, 70(2), 519–546.
- MENZEL, K. (2008): “Estimation and Inference with Many Moment Inequalities,” *Preprint, Massachusetts Institute of Technology*.
- MOON, H. R., AND F. SCHORFHEIDE (2009): “Bayesian and Frequentist Inference in Partially Identified Models,” *National Bureau of Economic Research Working Paper Series*, No. 14882.

- PAKES, A., J. PORTER, K. HO, AND J. ISHII (2006): “Moment Inequalities and Their Application,” .
- POLLARD, D. (1984): *Convergence of stochastic processes*. David Pollard.
- PONOMAREVA, M. (2010): “Inference in Models Defined by Conditional Moment Inequalities with Continuous Covariates,” .
- RIO, E. (1994): “Local invariance principles and their application to density estimation,” *Probability Theory and Related Fields*, 98(1), 21–45.
- ROMANO, J. P., AND A. M. SHAIKH (2008): “Inference for identifiable parameters in partially identified econometric models,” *Journal of Statistical Planning and Inference*, 138(9), 2786–2807.
- ROMANO, J. P., AND A. M. SHAIKH (2010): “Inference for the Identified Set in Partially Identified Econometric Models,” *Econometrica*, 78(1), 169–211.
- STONE, C. J. (1982): “Optimal Global Rates of Convergence for Nonparametric Regression,” *The Annals of Statistics*, 10(4), 1040–1053.
- STOYE, J. (2009): “More on Confidence Intervals for Partially Identified Parameters,” *Econometrica*, 77(4), 1299–1315.
- WICHURA, M. J. (1969): “Inequalities with Applications to the Weak Convergence of Random Processes with Multi-Dimensional Time Parameters,” *The Annals of Mathematical Statistics*, 40(2), 681–687.

Figure 1: 95% Confidence Region Using Weighted Sup Statistic (this paper)

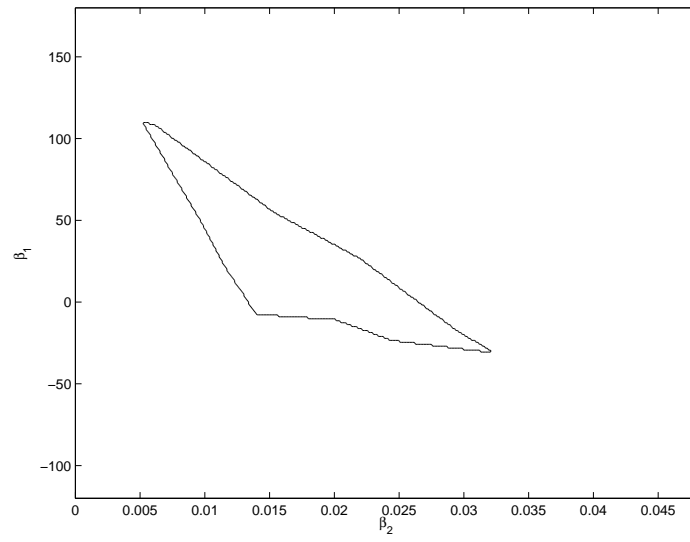


Figure 2: 95% Confidence Region Using Unweighted Statistic and Subsampling with Estimated Rate

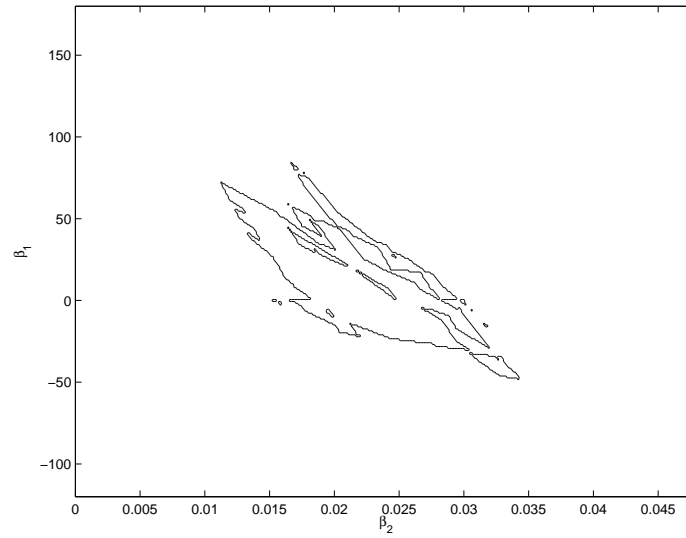
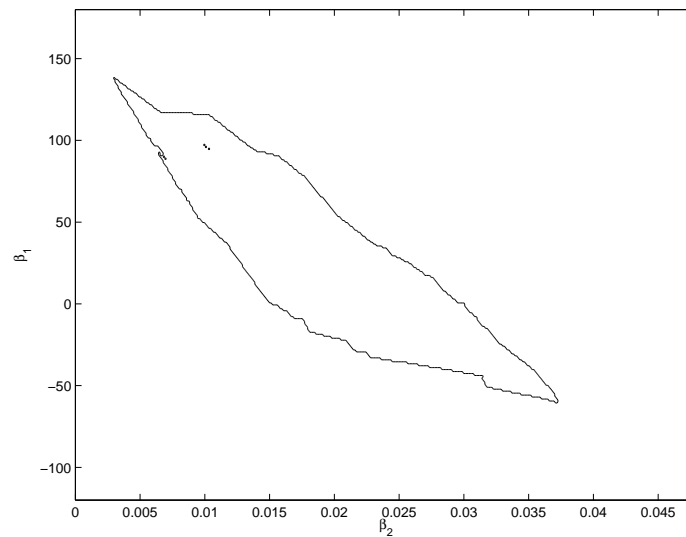


Figure 3: 95% Confidence Region Using Unweighted Statistic and Subsampling with Conservative Rate



	$t_n$	$n = 100$	$n = 500$	$n = 1000$
nominal size .1	$n^{-1/5}$	0.2510	0.1840	0.1770
	$n^{-1/3}$	0.1640	0.1160	0.1150
	$n^{-1/2}$	0.0890	0.0770	0.0880
nominal size .05	$n^{-1/5}$	0.1020	0.0650	0.0790
	$n^{-1/3}$	0.0750	0.0410	0.0550
	$n^{-1/2}$	0.0340	0.0220	0.0350

Table 1: False Rejection Probabilities for Least Favorable Null

$t_n$	$\theta_1 - \bar{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
$n^{-1/5}$	.1	0.2070	0.5030	0.7290
	.2	0.4800	0.9540	1.0000
	.3	0.7590	1.0000	1.0000
	.4	0.9560	1.0000	1.0000
	.5	0.9970	1.0000	1.0000
$n^{-1/3}$	.1	0.1440	0.4530	0.6300
	.2	0.3780	0.9390	0.9980
	.3	0.6910	1.0000	1.0000
	.4	0.8860	1.0000	1.0000
	.5	0.9820	1.0000	1.0000
$n^{-1/2}$	.1	0.1560	0.3580	0.5020
	.2	0.3480	0.8980	0.9910
	.3	0.6490	0.9990	1.0000
	.4	0.8620	1.0000	1.0000
	.5	0.9740	1.0000	1.0000

Table 2: Power for Level  $\alpha = .05$  Test with Critical Values Based on Finite Sample Least Favorable Distribution (Design 1)

$t_n$	$\theta_1 - \bar{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
$n^{-1/5}$	.1	0	0	0
	.2	0.0060	0.0160	0.0320
	.3	0.0260	0.1380	0.2950
	.4	0.0640	0.4490	0.8310
	.5	0.1750	0.8480	0.9950
$n^{-1/3}$	.1	0.0070	0.0120	0.0050
	.2	0.0160	0.0620	0.1000
	.3	0.0410	0.2150	0.4560
	.4	0.1190	0.6040	0.8760
	.5	0.2100	0.9020	0.9960
$n^{-1/2}$	.1	0.0060	0.0140	0.0100
	.2	0.0230	0.0570	0.0860
	.3	0.0380	0.2290	0.3890
	.4	0.1190	0.5320	0.7910
	.5	0.2030	0.8500	0.9820

Table 3: Power for Level  $\alpha = .05$  Test with Critical Values Based on Finite Sample Least Favorable Distribution (Design 2)

$t_n$	$\theta_1 - \bar{\theta}_1$	$n = 100$	$n = 500$	$n = 1000$
$n^{-1/5}$	.1	0.0340	0.0640	0.1200
	.2	0.0930	0.4660	0.7040
	.3	0.2720	0.8690	0.9900
	.4	0.5010	0.9940	1.0000
	.5	0.7670	1.0000	1.0000
$n^{-1/3}$	.1	0.0390	0.1040	0.1160
	.2	0.1120	0.4290	0.6400
	.3	0.2570	0.8380	0.9790
	.4	0.4630	0.9940	1.0000
	.5	0.7170	1.0000	1.0000
$n^{-1/2}$	.1	0.0300	0.0830	0.0870
	.2	0.1210	0.3250	0.5230
	.3	0.2400	0.7620	0.9670
	.4	0.3970	0.9840	1.0000
	.5	0.6690	1.0000	1.0000

Table 4: Power for Level  $\alpha = .05$  Test with Critical Values Based on Finite Sample Least Favorable Distribution (Design 3)

	$\theta_1$	$\theta_2$
Weighted Sup Statistic (this paper)	$[-30, 109]$	$[0.0053, 0.0320]$
Unweighted, Subsampling with Estimated Rate	$[-48, 84]$	$[0.0113, 0.0342]$
Unweighted, Subsampling with Conservative Rate	$[-60, 138]$	$[0.0030, 0.0372]$

Table 5: 95% Confidence Intervals for Components of  $\theta$