Essays on Set Estimation and Inference with Moment Inequalities

by

Konrad Menzel

Diplom, Universität Mannheim (2004)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2009

Author ....
Department of Economics
July 14, 2009

Certified by ....
Whitney K. Newey
Professor, Department of Economics
Thesis Supervisor

Certified by ....
Victor Chernozhukov
Professor, Department of Economics
Thesis Supervisor

Accepted by ....
Esther Duflo
Professor, Department of Economics
Chair, Department Committee on Graduate Theses

Essays on Set Estimation and Inference with Moment Inequalities

by

Konrad Menzel

Submitted to the Department of Economics
on July 14, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

This thesis explores power and consistency of estimation and inference procedures with moment inequalities, and applications of the moment inequality framework to estimation of frontiers in finance.

In the first chapter, I consider estimation of the identified set and inference on a partially identified parameter when the number of moment inequalities is large relative to sample size. Many applications in the recent literature on set estimation have this feature. Examples discussed in this paper include set-identified instrumental variables models, inference under conditional moment inequalities, and dynamic games. I show that GMM-type test statistics will often be poorly centered when the number of moment inequalities is large. My results establish consistency of the set estimator based on a Wald-type criterion, and I give conditions for uniformly valid inference under many weak moment asymptotics for both plug-in and subsampling procedures.

The second chapter evaluates the performance of an Anderson-Rubin (AR) type test for a finite number of moment inequalities, and propose a modified Lagrange Multiplier (LM) and a conditional minimum distance (CMD) statistic. The paper outlines a procedure to construct asymptotically valid critical values for both procedures. All three tests are robust to weak identification, however in most settings, conservative inference using the LM statistic seems to have greater power against local alternatives than the AR-type test. Furthermore, confidence regions based on the LM statistic will remain non-empty if the model is misspecified.

Finally, the third chapter, which is co-authored with Victor Chernozhukov and Emre Kocatulum, presents various set inference problems as they appear in finance and proposes practical and powerful inferential tools. Our tools will be applicable to any problem where the set of interest solves a system of smooth estimable inequalities, though we particularly focus on the following two problems: the admissible mean-variance sets of stochastic discount factors and the admissible mean-variance sets of asset portfolios. We propose to make inference on such sets using weighted likelihood-ratio and Wald type statistics, building upon and substantially enriching the available methods for inference on sets.

Thesis Supervisor: Whitney K. Newey
Title: Professor, Department of Economics

Thesis Supervisor: Victor Chernozhukov
Title: Professor, Department of Economics

Acknowledgments

This work would not have been possible without the constant support and encouragement from my advisors, Whitney Newey and Victor Chernozhukov. I am greatly indebted to them for many very helpful discussions over the past five years in which I learned how to be very rigorous in detail in my research while never losing out of sight the broader context and practical relevance of a theoretical question.

Jerry Hausman and Anna Mikusheva were also invariably very generous with their time and careful advice. I also owe thanks to other faculty at the department, especially Josh Angrist, Abhijit Banerjee, Esther Duflo, Panle Jia, and Stephen Ryan. I would also like to use this opportunity to express my deep gratitude to Axel Börsch-Supan, Bernd Fitzenberger, Stefan Hoderlein, Guido Imbens, and Konrad Stahl, who at various stages of my education encouraged and supported me in my plans to become a researcher in econometrics.

I also want to thank my office mates Chris Smith, Danielle Li, Jeanne Lafortune, Jim Berry, Leila Agha, Mandy Pallais, Neil Bhutta, Peter Hinrichs, and Tal Gross, and my fellow graduate students at the department, in particular Brigham Frandsen, Eric Weese, Greg Fisher, José Tessada, Mauro Alessandro, Paul Schrimpf, Raymond Guiteras, Rick Hornbeck, and Samuel Pienknagura for good company and many interesting discussions over the last five years. A special thanks goes to my two student coauthors Randall Lewis and Emre Kocatulum for enthusiastic and fruitful collaboration in our joint research projects.

I would also like to thank the staff at the economics department, in particular Jessica Colón, Emily Gallagher, Pete Hoagland, Gary King, and Katherine Swan, for assistance and advice on the many practical aspects of studying, teaching, and doing research at MIT. I acknowledge financial support by the German Ministry of the Economy through the ERP fellowship, and the Department of Economics at MIT.

Finally, I'd owe a deep gratitude to my family who accompanied me on this journey from the great distance and who have been a constant source of support to me over the last years.

Contents

List of Figures

Chapter 1

Estimation and Inference with Many Moment Inequalities

1.1 Introduction

In this paper, I consider estimation of the identified set and inference on a partially identified parameter when the number of moment inequalities is large relative to sample size. This situation is commonly found in applications in the fast-growing literature on partial identification. Prominent examples include estimation with conditional moment inequalities, instrumental variables models with missing or interval measured data, and estimation of games with rich strategy spaces. For instance Bajari, Benkard, and Levin (2007)'s procedure for the estimation of a dynamic oligopoly model uses up to 500 moment restrictions with a sample size of no more than 1,200 observations. Also, in point-identified problems, restricting attention to a subset of the available moment restrictions primarily affects only the efficiency of the estimator. However, in set estimation, omitting relevant constraints will also alter the shape of the identification region. Therefore in partially identified problems, estimation using a large number of moment restrictions is even more common than in the standard GMM framework.

In order to characterize the finite-sample properties of econometric procedures, I consider limits of sequences of experiments for which the number of moment inequalities grows at some rate as the sample size increases. In addition, the framework allows the combined strength of the moment conditions to change with sample size. The standard large-n asymptotic framework used in the previous literature implicitly constrains the number of moment conditions to be negligibly small and the identifying power of the moments to be proportional to sample size. The many weak moment approximations considered in this paper nest the standard setup as a special case, but also allow us to model more realistic settings in which identification is weak and the number of econometric restrictions is large.

This modification of the asymptotic experiment changes the conclusions of the previous literature in three main aspects:

(1) The distribution of the criterion used for estimation or inference need not degenerate on the interior of identification region, and standard test statistics need not be centered or attain their minimum in the identification region

(2) Standard approximations to the distribution of a vector of moment functions (subsampling, bootstrap, Gaussian) may be poor if its dimension is large, so that the true null rejection probabilities of hypothesis tests using critical values based on these approximations may exceed nominal size.

(3) Anderson-Rubin (AR)-type tests which are frequently used in the literature - including the Quasi-Likelihood Ratio (QLR) Statistic (Kudo (1963), Rosen (2008)) and the Empirical Likelihood Ratio (ELR) Statistic (Canay (2007)) - have many degrees of freedom. For moment equalities, the power of chi-square tests is known to decrease to size as the number of restriction goes to infinity, and furthermore for the one-sided testing problem, inference has to be conservative over high-dimensional nuisance parameter. Therefore tests based on these statistics should be expected to have low power if the number of moment restrictions is large.

The first point mainly concerns set estimation from lower contour sets of a GMM-type criterion, and will lead to inconsistency of the estimator unless severe restrictions on the dimension of the moment vector relative to the identifying power of the restrictions are imposed. The second aspect of the problem is relevant for construction of critical values for hypothesis tests and confidence sets and will be investigated in section 4 of this paper. The last observation concerns the choice of a test statistic and suggests that in many cases it will be possible to improve considerably over standard procedures by reducing the dimensionality of the parameter that is tested implicitly. As I will argue below, these three features of standard procedures will alter many of the recommendations put forward in the literature on set inference based on standard "large-n" asymptotics.

For point-identified problems, it has long been known that a large degree of over-identification often leads to significant finite-sample bias in GMM estimators and may render classical inference procedures invalid. GMM under weak identification with a fixed number of moments was considered in Stock and Wright (2000), and Han and Phillips (2006) analyze GMM with many weak moment conditions and give rates on the number of moments and their combined explanatory strength under which the GMM estimator is consistent and the GMM objective function converges to a non-stochastic limit. Newey and Windmeijer (2008) give conditions for consistency and derive the asymptotic theory for GMM, the Continuous Updating Estimator (CUE), and standard testing procedures under many weak moments sequences.

In this paper, I will argue that issues with finite-sample bias and bad size properties and power loss of common testing procedures typically associated with estimation with over-identification also arise in set-identified problems using many moment inequalities, which are in fact strictly under-identified according to conventional terminology. In many applications of set-identified models the relevant test statistic turns out to be minimized at a single point of the parameter space even if the parameter is only set-identified (e.g. Pakes, Porter, Ho, and Ishii (2006) and Bajari, Benkard, and Levin (2007)), and simulation studies often show substantial bias in the set estimator. This happens particularly often if the moment vector used for inference and estimation has a high dimension relative to sample size.

The problems of standard inference procedures are not necessarily limited to cases with a extremely large number of moments, but from the GMM literature it is known that finite-sample bias can be severe, even for a moderate degree of over-identification and especially if the identifying power of the moments is low. For example, Hansen, Heaton, and Yaron (1996) document significant bias of the 2-step GMM estimator for the CAPM for as few as 6 over-identifying restrictions with a sample of 400 observations.

In the literature on set-identified problems, consistency of criterion-based set estimators and validity of uniform confidence regions for the identified set based on the supremum of a GMM-type statistic on the identified set has been shown by Chernozhukov, Hong, and Tamer (2007) under standard "large-n" asymptotics. Inference on the true population parameter has been considered by Imbens and Manski (2004), Chernozhukov, Hong, and Tamer (2007), and Andrews and Guggenberger (2007b).

In the theoretical literature, set inference subject to infinitely many moment restrictions has only been considered systematically by Chernozhukov, Lee, and Rosen (2008), Kim (2008), and Andrews and Shi (2008). Andrews and Guggenberger (2007b) give conditions for uniformly valid inference for a fixed number of moment conditions under local parameter sequences which include cases in which some moment inequalities are close to binding. This covers in particular the set-identified analogue of the problem of weak identification.

The primary contribution of this paper is to analyze commonly used procedures for set estimation and inference under many moment asymptotics. I derive conditions on the number of moment restrictions used for estimation and the combined explanatory power of those restrictions under which different estimators of the identified set are consistent. I discuss these conditions for a number of practically relevant examples. I also find that for a slow to intermediate growth rate m_n for the number of moments, critical values for GMM-type statistics based on a normal approximation yield uniformly valid inference, whereas subsampling critical values are valid only for slow rates in m_n. The reason for the poor performance of subsampling is that subsampling will in general fail to approximate distributional features of the moment vector other than only the first two moments. In situations in which the number of moments is large relative to sample size, the resulting critical values need not even be conservative, but fail to guarantee the desired confidence level altogether.

As an example, I develop an inference procedure for conditional moment inequalities based on series approximations. For the case of a one-dimensional conditioning set, I show that if the number of unconditional moments is chosen as to achieve the fastest possible rate of convergence for the corresponding set estimator, Gaussian asymptotic approximations to the distribution of any of the commonly used test statistics discussed in section 4 continue to be valid.

This paper proceeds as follows: In section 2, I will outline the problem and give basic notation. Section 3 analyzes the behavior of GMM-type criterion functions under many weak moments asymptotics and gives conditions for consistency of set estimators defined as lower contour sets of the criterion. Section 4 gives conditions for uniformly valid set inference under many moment sequences with drifting parameters. Section 5 concludes.

1.2 Setup

In this paper, I consider inference on a k-dimensional parameter $\theta \in \Theta$ given a sample $Y_{1n}, \dots, Y_{nn}$ of n observations. The observed sample is modeled as a triangular array of random variables $Y_{1n}, \dots, Y_{nn}$ which are i.i.d. from a population distribution $P_n \in \mathcal{P}$ for each n.[1] It is possible to relax the i.i.d. assumption, but for expositional purposes, I will only consider the leading case of i.i.d. observations in this paper.

Estimation and inference will be based on an m-dimensional vector $g_m(Y_i, \theta)$, where the population parameter θ_0 satisfies

$$\mathbb{E}_{P_n}[g_m(Y_{in}, \theta_0)] \geq 0 \tag{1.1}$$

for all $P_n \in \mathcal{P}$. I will allow $m = m_n$ to increase at a certain rate as the sample size grows. Throughout the paper, I will treat the order at which additional moment

[1] Following the notation in van der Vaart and Wellner (1996), P_n will always represent the population distribution for the nth row vector, whereas the empirical measure will be denoted $\mathcal{P}_n$.

inequalities are imposed as fixed, and state asymptotic results depending only on the rate m_n at which new moments are added.

1.2.1 Examples

There are many econometric problems in which the number of moment inequalities can be very large. As a first example, we consider a linear model which allows for a large number of unconditional moment restrictions, and which is similar in spirit to Manski and Pepper (2000)'s "Monotone Instrumental Variable" (MIV) setting.

Example 1 Linear "One-Sided" Instrumental Variables *Suppose we have variables* Z_{im} *which do not satisfy a proper exclusion restriction in a regression of* Y_i *on* X_i*, but we know sign of bias. The moment restrictions are of the form*

$$g_l(\theta, P) := \mathbb{E}[Z_{il}(Y_i - X_i\theta)] \geq 0 \text{ for } l = 1, \ldots, m$$

An estimation problem with this structure can arise in many situations, e.g.

- *differential sample attrition*
- *with heterogeneity in parameters, want to bound one particular average treatment effect with local average effects*
- *Manski and Pepper (2000)'s Monotone IV assumption*
- *identification from discrete variation (Chesher (2005))*

Generally, the number of instruments in this setting can be large for the same reasons as in point-identified settings.

A variation of this example would be IV regression with interval-measured data, which is related to the problem analyzed by Manski and Tamer (2002) and has been analyzed by Bontemps, Magnac, and Maurin (2007).

Another important case in which the number of moment functions is potentially infinite is that of conditional moment restrictions. This arises frequently, for example in structural estimation with instrumental variables. Examples include Manski and Tamer (2002)'s framework for estimation of bounds for linear models with interval-measured data or Khan and Tamer (2008)'s estimation of censored regression models. Also in the setup of Pakes, Porter, Ho, and Ishii (2006) in the estimation of games with incomplete information, any quantities that are common knowledge among all players and observed by the econometrician can be used as instrumental variables.

Example 2 Conditional Moment Restrictions I *Suppose for an i.i.d. sample of observations $W_i = (Z_i, Y_i)$ we have a moment restriction of the form*

$$h(z, \theta_0, P) := \mathbb{E}[\varrho(Y_i, \theta_0)|Z_i = z] \geq 0 \quad \text{for all } z \in \mathcal{Z} := \text{supp}G(z)$$

where $Z_i \sim G(z)$ is a vector of instrumental variables with a continuous distribution, and we assume for simplicity that $\mathcal{Z}$ is bounded and the density of Z_i is bounded away from zero uniformly on $\mathcal{Z}$. The residual $\varrho(Y_i, \theta)$ is a real-valued function of the data W_i and a parameter vector θ.[2]

We can now form moment functions $g_l(W_i, \theta, P) := \psi_l(Z_i)\varrho(Y_i, \theta)$ for a given choice of non-negative instruments $\psi_m(Z_i)$, so that at the population parameter θ_0,

$$\mathbb{E}_P\left[g_l(W_i, \theta_0)\right] := \mathbb{E}_P[\psi_l(Z_i)\varrho(Y_i, \theta_0)] = \mathbb{E}_P\left[\psi_l(Z_i)h(Z_i, \theta_0, P)\right] \geq 0$$

by the law of iterated expectations. As we will discuss below, possible choices of instruments include basis functions for B-Spline approximations (see e.g. Nürnberger (1989)), or characteristic functions for subintervals of $\mathcal{Z}$ as in Andrews and Shi (2008).

For expositional purposes, I will now propose an alternative way of forming unconditional moments from the conditional moment inequality model which is better suited for the subsequent discussion of the rates of consistency of set estimators.

[2] This can be generalized easily to a vector valued residual function, but for notational simplicity, we will stick to the one-dimensional case.

Example 3 Conditional Moment Restrictions II *Consider the conditional moment inequality model from Example 2. If $h(z,\theta,P)$ is continuous in z for any value of θ, we can approximate the function using B-splines.*[3] *Given a matrix $\Psi_m := \{\psi_l^m(Z_i)\}_{i=1,l=1}^{i=n,l=m}$ of m basis functions $\psi^m(z) := (\psi_1^m(z),\dots,\psi_m^m(z))'$, we have*

$$h(z,\theta,P) = \sum_{l=1}^{m} \psi_l^m(z)\pi_l^m(\theta,P) + R_m(z,\theta,P) = \psi^m(z)'\pi^m(\theta) + R_m(z,\theta)$$

for some remainder term $R_m(z,\theta)$ such that $\int R_m(z,\theta)^2 dG(z)$ is minimized, i.e. results from a projection of $h(z,\theta,P)$ onto the spline space generated by $\psi^m(z)$ with respect to the weighted L_2 norm, where the weights are given by the distribution of Z_i. It is known that any nonnegative function can be approximated by B-splines with nonnegative coefficients (see De Boor and Daniel (1974)), so that we can consider a (possibly data-dependent) restricted projection of $h(z,\theta,P)$ onto the spline space with positive coefficients.

For example, we could seek to minimize the length of $Q_{\Psi_m}(h(Z_n,\theta,P_n) - \Psi_m\pi)$, where $Q_{\Psi_m} = \Psi_m(\Psi_m'\Psi_m)^-\Psi_m'$ is the linear projector onto the column space of Ψ_m,[4] *and A^- denotes the generalized (Moore-Penrose) inverse of a square matrix A. Then this amounts to solving*

$$\min_{\pi\geq 0} \|Q_{\Psi_m}(r_n - \Psi_m\pi)\| = \min_{t\geq 0}(\hat{\pi}^m - t)'(\Psi_m'\Psi_m)(\hat{\pi}^m - t)$$

where $r_n(\theta) := (\varrho(Y_1,\theta),\dots,\varrho(Y_n,\theta))'$. Hence, a test based on unconditional moments that are formed using instruments $\psi^m(Z_i)$ can be interpreted as testing whether the least-squares coefficients $\hat{\pi}_m := (\Psi_m'\Psi_m)^-\Psi_m' r_n$ from the unrestricted projection of r_n onto the basis functions of the spline space are non-negative. A test of this form

[3] For m equidistant knots $t_1 < \cdots < t_l < \cdots < t_m$, the basis B-Spline of order n can be constructed recursively as $\psi_{l,n}(t) := \frac{t-t_l}{t_{l+n}-t_l}\psi_{l,n-1}(t) + \frac{t_{l+n+1}-t}{t_{l+n+1}-t_{l+1}}\psi_{l+1,n-1}(t)$, where we set $\psi_{l,0}(t)$ equal to the characteristic function for the interval $[t_l, t_{l+1})$.

[4] Note that by the definition in footnote 3, $\psi_{l,n}(t)$ has support only on the interval $[t_l, t_{l+1})$, so that if the p.d.f. of Z is bounded from below on $\mathcal{Z}$, and the uniform partition $t_1 < \cdots < t_m$ grows finer at a rate slower than n^{-1}, the smallest eigenvalue of the matrix $\Psi_m'\Psi_m$ will be bounded from below by a positive constant with probability going to 1. For the remaining discussion of this example, we can assume for simplicity a fixed design setting with regard to the values $Z_i = Z_{in}$, where, without loss of generality, the draws of Z_{in} are evenly spaced on $\mathcal{Z}$.

clearly has power against any alternative θ_A because if $h(z, \theta_A, P) < 0$ at some value of z, in the limit at least one spline coefficient has to be negative.[5]

Conditional moment inequalities are a special case of set estimation subject to a continuum of inequality constraints, which has been analyzed for the case of intersection bounds for a one-dimensional parameter θ by Chernozhukov, Lee, and Rosen (2008) who propose both kernel and series based methods to construct implied bounds on the parameter.

Finally, we consider moment inequalities from economic models of optimization behavior and estimation of discrete games:

Example 4 Estimation of Discrete Games *Suppose a group of n agents can make a choice $s \in \mathbb{S}$, where $\mathbb{S} = \{s_1, \dots, s_m\}$ is a finite set of pure strategies common to all agents. The information set of the agent is given by the variables Z_i, and we observe the agent's choice S_i as well as her opponents' strategy profile S_{-i}. Therefore for the population parameter θ_0 we have*

$$h(z, \theta_0, P) := \mathbb{E}_P[\pi(Y_i, S_i, S_{-i}, \theta_0) - \pi(Y_i, s', S_{-i}, \theta_0)|Z_i = z] \geq 0 \quad \forall s' \in \mathbb{S}$$

Hence for each $s' \in \mathbb{S}$, we can form moment conditions

$$g_m(Z_i, Y_i, \theta) = \psi(Z_i) \otimes \begin{bmatrix} \pi(Y_i, S_i, S_{-i}\theta) - \pi(Y_i, s_1, S_{-i}, \theta) \\ \vdots \\ \pi(Y_i, S_i, S_{-i}\theta) - \pi(Y_i, s_m, S_{-i}, \theta) \end{bmatrix}$$

where $\psi(z)$ is a vector-valued positive function of the conditioning variable. The dimension of the moment vector $g_i(\theta)$ can be large if either the strategy space $\mathbb{S}$ or the information set is very rich.

[5] This idea extends to the general case of a continuum of moment conditions in a straightforward manner. For example in the oligopoly model in Bajari, Benkard, and Levin (2007), investment σ is a continuous strategy, so it would be possible to replace the vector r_n with simulated payoff differentials for appropriately chosen values of σ and let Ψ_m be a matrix of B-spline basis functions in σ. This method would aggregate the information from a large number of values for σ to a moment vector whose dimension is lower by an order of magnitude and should be chosen depending on sample size.

Symmetry and discreteness of the game are imposed only for notational convenience, and Pakes, Porter, Ho, and Ishii (2006) also discuss extensions if the information set is not common knowledge among the agents and the econometrician. In the estimation of a Dynamic Oligopoly Model, Bajari, Benkard, and Levin (2007), firms' strategies are assumed to be stationary but depend on a rich state space and entail both discrete entry/exit and continuous investment decisions, so that there is a large number of alternative strategies. In their example, for a sample of at most $n = 1600$ observations Bajari, Benkard, and Levin (2007) draw as many as $m = 500$ alternative strategies at random and construct moments from differences in instantaneous profits and simulated value functions.

1.2.2 Identification

For a fixed sample size n, the identification region $\Theta_{I,n}$ is defined as the subset of Θ for which the population moment restrictions in (1.1) hold,

$$\Theta_{I,n} := \{\theta \in \Theta : \mathbb{E}_{P_n}[g_m(Y_{in}, \theta_0)] \geq 0\}$$

The second subscript indicates that the identification region will be allowed to change with sample size both through the population distribution P_n of Y_{in}, and the number m_n of moment inequalities imposed for estimation or inference. Note that indexing the identification region with sample size n is not meant to suggest a dependence on the particular realization of the sample. Also, even though the econometric model is incomplete in the sense that the moment conditions (1.1) hold at θ_0 for every measure $P \in \mathcal{P}$, the identification region is defined with respect to one particular population measure $P_n \in \mathcal{P}$.

For a large, potentially infinite, number of moment inequalities, the main object of interest for estimation is the set of parameter values that satisfy all moment restrictions that can be derived from the econometric model. The next section will give a formal definition of the *sharp identification region* Θ_I as the (set-valued) limit of the approximating sequence $\Theta_{I,n}$.

The identification region can be characterized as the (typically set-valued) arg-zero of the population criterion

$$Q_n(\theta) := \min_{t\geq 0} \left(\mathbb{E}_{P_n}[g_M(Y_{in},\theta_0)] - t\right)' W_n(\theta) \left(\mathbb{E}_{P_n}[g_m(Y_{in},\theta_0)] - t\right) \tag{1.2}$$

where the $m_n \times m_n$ matrix $W_n(\theta)$ is continuous and positive definite. For a given value of θ, the minimizer t^* of the quadratic form over non-negative values of $t \in \mathbb{R}^M_+$ is the projection of the moment vector onto the positive orthant with respect to the Euclidean norm defined by $W_n(\theta)$. Loosely speaking, concentrating out the slackness parameter $t \in \mathbb{R}^m_+$ can be understood as penalizing only the component-wise negative parts of the moment vector. Under conditions to be discussed in the next section, the criterion $Q_n(\theta)$ defined in (1.2) epi-converges to a limit $Q_0(\theta)$, where $Q_0(\theta) = 0$ if and only if $\theta \in \Theta_I$, the sharp identification region.

One particular case of interest is that of a continuum of moment conditions of the form

$$h(z,\theta_0,P) \geq 0 \text{ for each } z \in \mathcal{Z} \text{ and } P \in \mathcal{P}$$

where $\mathbf{h}(\theta,P) := (h(z,\theta,P))_{z\in\mathcal{Z}}$ is a vector in the separable Hilbert space $L_2(\mathcal{Z},P)$ which depends on the probability measure P. This moment vector does not necessarily have to be an unconditional expectation of a known function of the data, but e.g. in the case of a conditional moment inequality as in Example 2, $h(z,\theta,P)$ is the conditional expectation function $\mathbb{E}_P[\varrho(Y_i,\theta)|Z_i = z]$. The corresponding population criterion function is

$$Q_0(\theta) := \inf_{\varphi\geq 0} \int_{\mathcal{Z}\times\mathcal{Z}} (h(s,\theta,P_n)-\varphi(s))(h(z,\theta,P_n)-\varphi(z))w_{n,\theta}(s,z)P_n(ds)P_n(dz) \tag{1.3}$$

for a positive definite weighting kernel $w_\theta(s,z)$.[6] Since $\mathbb{R}^M_+$ is a closed convex subset

[6] For example, the criterion based on approximation of the continuum through averages on subintervals,

$$Q_0(\theta) = \min_{\varphi\geq 0} \int_{\mathcal{I}} (g(I) - \varphi(I))'W(I)(g(I) - \varphi(I))\mu(dI)$$

given a weighting distribution $\mu(I)$ over the set $\mathcal{I}$ of subintervals $I \subset \mathcal{Z}$, where $g(I) := \frac{\int_I h(z)P(dz)}{\int_I P(dz)}$. For simplicity, suppose the weighting matrix is not data-dependent and diagonal with weight $\omega(\mathcal{I})$

of a Hilbert space, the infimum $t^* = t^*(\theta, P)$ in the definition (1.3) is attained and the arginf is unique (see e.g. section 3.12, Theorem 1 in Luenberger (1969)). For estimation, $Q_0(\theta)$ can be approximated using a finite-dimensional vector of unconditional moments.

1.2.3 General Approach to Estimation and Inference

For inference we will replace population moments with their sample counterparts

$$\mathbb{E}_n[g(Y_{in}, \theta)] = \hat{g}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} g_i(\theta)$$

where the covariance matrix of the moment vector is given by $\Omega(\theta) = \mathrm{Var}\left(\sqrt{n}\hat{g}_n(\theta)\right)$. Using the sample moments, we can form the sample criterion

$$\hat{Q}_n(\theta) := \min_{t\geq 0}(\hat{g}_n(\theta) - t)'W_n(\theta)(\hat{g}_n(\theta) - t) \tag{1.4}$$

for a potentially data-dependent weighting matrix $\hat{W}_n(\theta)$ that converges to $W_n(\theta)$ uniformly in the sense that $\sup_{1\leq k,l\leq m_n}\left|\hat{W}_{n,kl}(\theta) - W_{n,kl}(\theta)\right| \xrightarrow{p} 0$ as $n \to \infty$, where A_{kl} denotes the (k,l) element of a matrix A.

Given the sample criterion from (1.4), we can construct the set estimator as a lower contour set of $\hat{Q}_n(\theta)$ for a data-dependent non-negative sequence $\hat{c}_n$,

$$\hat{C}_n := \left\{\theta \in \Theta : n\hat{Q}_n(\theta) \leq \hat{c}_n\right\} \tag{1.5}$$

on interval $\mathcal{I}$. Then the criterion function can be represented using the kernel

$$w(s,z) := \int_{\mathcal{I}} W(I)\mathbb{1}\{s \in I, z \in I\}\mu(dI)$$

For example if $\mathcal{Z} \subset \mathcal{R}$ is an interval of length $\Delta = |\bar{z} - \underline{z}| > 1$, and intervals are drawn from a distribution that puts mass proportional to ω_p on any subinterval of length 2^{-p}, $p = 0, 1, 2, \ldots$, the implicit kernel is proportional to

$$w(s,z) \propto \sum_{p=1}^{\infty}\omega_p \max\left\{\frac{2^{-p} - |s-z|}{\Delta - 2^{-p}}, 0\right\}$$

The next section is going to discuss conditions on $\hat{c}_n$ which ensure consistency of the set estimator for Θ_I. Alternatively, as proposed by Chernozhukov, Hong, and Tamer (2007), we can construct a sequence $\hat{c}_n$ of cutoff values such that $\hat{C}_n$ is a valid $1-\alpha$ confidence set either for the population parameter θ_0 or the sharp identification region Θ_I.

1.2.4 Comparison with Conditional Moment Equalities

The consistency results in section 3 imply that with an infinite number of moment inequalities, there will typically be no estimator of the form (1.5) that is $\sqrt{n}$-consistent for the sharp identification region Θ_I. This contrasts with well-known convergence results for the point-identified setting with infinitely many moment equalities analyzed among others in Newey (1990), Carrasco and Florens (2000), Ai and Chen (2003), and Domínguez and Lobato (2004), and I am going to devote the remainder of this section to give an intuitive explanation for this difference.

In the literature on estimation subject to a continuum of moment conditions, consistency is usually achieved imposing a full-rank condition on the Hessian of $Q_0(\theta)$,[7]

$$\nabla^2_{\theta\theta} Q_0(\theta_0) = \int \nabla_\theta h(s, \theta_0, P) \nabla_{\theta'} h(z, \theta_0, P) w_{\theta_0}(s, z) P(ds) P(dz)$$

on the boundary of the identified set Θ_I.[8] If this condition fails, the error in the nonparametric estimation of the moment functions $h(z, \theta, P)$ may dominate in the limit in the approach of Newey (1990) and Ai and Chen (2003) and slow down the rate of consistency of the point estimator.

[7] The following condition corresponds to Assumption 4.1. in Ai and Chen (2003), and is implicit in the statement of Theorem 2 in Domínguez and Lobato (2004). Generally speaking, in order to achieve consistency of the estimator, the infimum of the population criterion $Q_0(\theta)$ which is by definition achieved at the identification region Θ_{In}, has to be well-separated (see e.g. van der Vaart and Wellner (1996))

$$\inf_{\theta \in S_n(\delta)} Q_0(\theta) \geq \kappa\delta^2 \tag{1.6}$$

for any $\delta > 0$, where $\kappa > 0$ and $S_n(\delta) := \{\theta \in \Theta : \delta/2 < d(\theta, \Theta_{In}) < \delta)$. This statement will be made more precise in the analysis of consistency for the set estimator in section 3 of this paper.

[8] Note that even though the kernel is also allowed to depend on θ, it is straightforward to verify that the additional derivative terms have expectation equal to zero for any value of θ in the identified set.

Since only the part of the continuum corresponding to binding constraints will contribute to the Hessian, a necessary condition for the full-rank condition to hold is that for any θ_0 on the boundary of the identified set, the binding restrictions constitute a subset of the continuum with measure bounded away from zero. In the point-identified case, this assumption is very natural since in this case, the identification region consists of one unique parameter value θ_0 satisfying the moment condition P-a.s., so that a mass of the continuum with strictly positive measure must be binding at θ_0. This need in general not be the case for set-identified models: even if for any parameter value $\theta \in \Theta_I^C$ outside the identification region, $\mu(\theta) := P(h(Z, \theta, P) \leq 0) > 0$, under reasonable conditions $\inf_{\theta \in \Theta_I^C} \mu(\theta) = 0$.

Except in some very special cases, the moment condition will be slack for all z except at a point z^* at any $\theta_0 \in \partial\Theta_I$, so that if P is absolutely continuous with respect to Lebesgue measure, the second derivative of $Q_0(\theta)$ is defined,[9] and

$$\nabla^2_{\theta\theta'} Q_0(\theta) := \int \nabla_\theta h(s, \theta, P) \nabla_{\theta'} h(z, \theta, P) \mathbb{1}_{\{h(s,\theta,P)=h(z,\theta,P)=0\}} w_\theta(s, z) P(ds) P(dz) = 0$$

Hence the analogue of the rank condition driving the $\sqrt{n}$ consistency results for estimation subject to a continuum of moment equalities will likely fail in set-identified problems with infinitely many moment inequalities.

This also has implications for power of tests against local alternatives: since the subset of the continuum of violated constraints will typically shrink as the parameter sequence $\theta_n = \theta_0 + \lambda_n \in \Theta_I^C$ approaches a point θ_0 on the boundary of the identification region, and in that case, the population criterion $Q_0(\theta)$ will vanish at a rate faster than $O((\theta_n - \theta_0)^2)$.[10]

[9] If at some $\theta_0 \in \partial\Theta_I$, the measure of the continuum $\mu(\theta_0)$ corresponding to binding conditions is strictly positive, the criterion defined in 1.3 will not be differentiable at θ_0, but the set of second subdifferentials will be defined (for definitions, see Rockafellar and Wets (1998)) and may contain a non-zero element.

[10] I.e. for many realistic settings, a condition like Kim (2008)'s Assumption 4.1(g) that the criterion is locally quadratic in the distance to the identified set seems to restrict the estimation problem to cases in which the identified set is defined by a finite subset of the moment inequalities: suppose e.g. that $\mathcal{Z}$ is bounded, and the density of Z is bounded from above by $\bar{p}$. Furthermore, let $h^*(z, \theta, P)$ be quasi-concave in z, and there is $B > 0$ such that for any $\theta \in \Theta_I^C$, $P(h^*(Z, \theta, P) > 0) > 0$. Furthermore Then for $\varepsilon := \frac{B}{4\bar{p}} > 0$, there is a finite ε-packing set of points $\bar{\mathcal{Z}} = \{z_1, \dots, z_M\}$, such that for each $\theta \in \Theta_I^C$, $h^*(z_i, \theta, P) > 0$ for at least one value of $\bar{\mathcal{Z}}$. Hence we could construct a

1.3 Consistency of the Set Estimator

In this section, I will consider set estimators obtained from inverting a criterion function at a possibly data-dependent cutoff value that depends on sample size but is fixed across parameter values. This type of set estimators correspond to fixed critical value confidence sets for the identified set as those proposed by Chernozhukov, Hong, and Tamer (2007) or Romano and Shaikh (2006) where I let the confidence size shrink to zero at some rate as the sample size increases.

In order to define the sharp identification region Θ_I when the number of moment restrictions increases in sample size, I start from the identified set $\Theta_{I,n}$ for a finite subset of the moment restrictions, and then take the limit as I let the number of restrictions used for inference go to infinity. I therefore first have to introduce the notion of Painlevé-Kuratowski set convergence (see also Molchanov (2005) or Rockafellar and Wets (1998)):

Definition 1 *For a sequence A_n of sets, the* inner limit $\liminf A_n$ *is the collection of the limit points x for which we can construct a converging sequence $x_n \rightarrow x$ such that $x_n \in A_n$ for all n. The* outer limit $\limsup A_n$ *is the set of points x for which we can construct a converging subsequence $x_{n(k)} \in A_{n(k)}$ such that $x_{n(k)} \rightarrow x$. We say that A_n* PK-converges *to A, in symbols $\lim_n A_n \overset{PK}{\rightarrow} A$, if $\liminf_n A_n = \limsup_n A_n = A$.*

Alternatively, the inner limit contains all points which are attainable through a sequence such that $x_n \in A_n$ for all except finitely many values of $n \geq 1$, whereas the outer limit consists of the limit points of sequences for which $x_n \in A_n$ for infinitely many $n \geq 1$. In this paper, I will only consider the case in which the parameter space Θ is a bounded subset of $\mathbb{R}^k$. Under this assumption, PK set convergence for nonempty closed sets is metrized by the Hausdorff distance of two sets A and B,

$$d_H(A, B) = \max\left\{\sup_{a\in A} d(a, B), \sup_{b\in B}(A, b)\right\}$$

finite-dimensional sieve space that is identified with a finite set $\bar{Z}$ of evenly-spaced values $z \in \mathcal{Z}$, and which contains the identified set.

Condition 1 (Identified Set) *(a) The parameter space $\Theta \subset \mathbb{R}^k$ is nonempty and compact. (b) The identified set is given by $\Theta_{I,n} = \{\theta \in \Theta : \mathbb{E}_{P_n}[g(Y_i, \theta)] \geq 0\}$, and $\Theta_I = \Theta_{I,\infty} := \lim_n \Theta_{I,n}$ in the sense of Painlevé-Kuratowski set convergence.*

Note also that if $\bar{g}_n(\theta)$ is continuous in θ and the sign of $\bar{g}_n(\theta)$ doesn't change in n for any value of θ, PK convergence Condition 1 (b) is satisfied, and we have

$$\Theta_I = \bigcap_{n \geq 0} \Theta_{I,n}$$

since by definition, the sequence $\Theta_{I,n}$ is nonincreasing in n with respect to the partial ordering induced by set inclusion, $\subset$ and

$$\liminf_n \Theta_{I,n} = \limsup_n \Theta_{I,n} = \bigcap_{n \geq 0} \mathrm{cl}\Theta_{I,n} = \bigcap_{n \geq 0} \Theta_{I,n}$$

by a straightforward argument, so that indeed $\Theta_I = \bigcap_{n \geq 0} \Theta_{I,n}$. Note however that $\Theta_{I,n}$ need not necessarily be nonincreasing in n, as example 5 below illustrates.

It is now useful to indicate the speed at which the identified set at sample size n converges to its limit under the Hausdorff metric:

Condition 2 *There is a non-increasing sequence τ_n of non-negative constants such that*

$$d_H(\Theta_{I,n}, \Theta_I) = O(\tau_n)$$

It will usually not be straightforward to derive the rate τ_n from the primitives of the problem, but we can continue the discussion of the conditional moment inequality problem in Example 3 to illustrate the approximation property of $\Theta_{I,n}$ with respect to the sharp identification region.

Example 5 *Consider the setting of Example 3. Suppose z is scalar, and for every $n = 1, 2, \dots$, $h(z, \theta, P_n)$ has bounded second derivatives in (z, θ) at all values of $z \in \mathcal{Z}$, where for now I assume that $\mathcal{Z}$ is a compact subset of $\mathbb{R}$. Also let $D_n(z, \theta) := \frac{\partial}{\partial \theta} h(z, \theta, P_n)$, and suppose that there is a sequence a_n of constants such* *that $a_n^{-1/r} n^{1/2} \|D_n(Z_i, \theta)\|$ is bounded away from zero. Also assume that for some sequence $p_m \to 0$, $\sup_{z \in \mathcal{Z}} \|[h(z, \theta, P_n) \vee (-\kappa)] - [\psi^m(z)'\pi(\theta)^m \vee (-\kappa)]\| = O(c_m)$ for some $\kappa > 0$ and any $\theta \in \Theta$ (e.g. by Proposition 2.8 in De Boor and Daniel (1974), for B-splines with nonnegative coefficients of fixed order k with m evenly spaced knots, $c_m = m^{-2}$). Then, as shown in the appendix, Condition 2 holds with $\tau_n = \frac{c_{m_n} n^{1/2}}{a_n^{1/r}}$.*

In order to allow for the rate of convergence of the set estimator to differ across the k dimensions of the parameter space, we are now going to define a rescaled Hausdorff distance. For some deterministic sequence $\mu_{1n}(\theta), \dots, \mu_{kn}(\theta)$ of appropriately chosen constants and some positive number $r > 0$ (all of which will be determined by Condition 4 below) I specify a parameter-dependent diagonal matrix

$$S_{n,\theta} = \operatorname{diag}(\mu_{1n}(\theta)^{1/r}, \dots, \mu_{kn}(\theta)^{1/r}) \tag{1.7}$$

In the following, I am going to use the properly renormalized Hausdorff-metric

$$\varrho_n(A, B) := \min\left\{\sup_{\theta \in A} d(S_{n,\theta}\theta, S_{n,\theta}B), \sup_{\theta \in B} d(S_{n,\theta}A, S_{n,\theta}\theta)\right\} \tag{1.8}$$

which differs from the usual Hausdorff distance in that the scaling of the local parameter space inside the supremum depends on the order of arguments. In a slight abuse of notation, I will also denote the pseudo-distance of a point from a set by $\varrho_n(\theta, A) := \varrho_n(\{\theta\}, A)$.

I will now develop an abstract consistency result for set estimation in terms of the sample criterion function $\hat{Q}_n(\theta)$, which will then be applied to the set estimation problem outlined in section 2. The criterion function $n\hat{Q}_n(\theta)$ can be decomposed into

$$n\hat{Q}_n(\theta) = \mu_n \gamma_n(\theta) + m_n \delta_n(\theta) + R_n(\theta)$$

where $\gamma_n(\theta)$ and $\delta_n(\theta)$ are non-stochastic functions. For the moment inequality model, $\gamma_n(\theta)$ will have the interpretation of the identifying ("signal") content of the population moments, and $\delta_n(\theta)$ will be the expectation of the "noise" contribution

$\zeta_n(\theta) := \hat{g}_n(\theta) - \mathbb{E}_{P_n}[g(Y_i,\theta)]$ of the sample moments. I will also define $\alpha_n := \frac{m_n}{\mu_n}$ where the focus of attention will be on cases in which $\alpha_n \to \alpha \in [0,\infty)$ and $\mu_n \to \infty$ as $n \to \infty$.

In order to allow the strength of identification in terms of asymptotic rates to differ over the parameter space, I have to scale the criterion in a way which may potentially result in it or its components taking infinite values, a case which has been considered in the literature on constrained M-estimation among others by Geyer (1994) and Knight (1999). Under these conditions, we can typically not achieve uniform convergence, but I will rely on the weaker notion of epi-convergence (see e.g. Rockafellar and Wets (1998)): Recall that a function $f(\theta)$ is lower semi-continuous (l.s.c.) if for any sequence $\theta_n \to \theta_0$, $\liminf_n f(\theta_n) \geq f(\theta_0)$. We then say that a sequence $f_n(\theta)$ of l.s.c. functions epi-converges to a l.s.c. function $f(\theta)$, $f_n(\theta) \overset{\text{epi}}{\to} f(\theta)$, if for every sequence $\theta_n \to \theta_0$ one has $\liminf_n f_n(\theta_n) \geq f(\theta_0)$, and there is some sequence $\theta_n \to \theta_0$ such that $\limsup_n f_n(\theta_n) \leq f(\theta_0)$.[11]
We can now state our main conditions on the criterion function:

Condition 3 (Criterion Function) *The criterion function $n\hat{Q}_n(\theta)$ is nonnegative and lower semi-continuous and*

(a) The rescaled population criterion function $\gamma_n(\theta)$ is nonnegative and lower semi-continuous, $\arg\inf_\theta \gamma_n(\theta) = \Theta_{I,n} \subset \Theta$, *and* $\inf_\theta \gamma_n(\theta) = 0$.

(b) $\gamma_n(\theta) \overset{\text{epi}}{\to} \gamma(\theta)$.

(c) For some constant $0 < K < \infty$,

$$\sup_{\theta\in\Theta} \left| \min\{K, \mu_n^{-1} n\hat{Q}_n(\theta) - \alpha_n \delta_n(\theta)\} - \min\{K, \gamma_n(\theta)\} \right| \xrightarrow{p} 0$$

(d) $\delta_n(\theta)$ is uniformly bounded in $\theta \in \Theta$

[11] Note that this condition is equivalent to convergence of the epi-graphs $\text{epi} f_n := \{(\theta, y) : y \geq f_n(\theta), \theta \in \Theta\}$ to $\text{epi} f$ with respect to PK set convergence, see Rockafellar and Wets (1998) Proposition 7.2.

Part (a) is mainly needed to ensure that the identification region as defined through the population criterion function is closed, and that the limit in part (b) yields a well-defined minimization problem whose solution will correspond to the sharp identification region.[12] Part (c) requires uniform convergence in probability, where the truncation at a fixed level K avoids problems in cases for which $\gamma_n(\theta)$ diverges to infinity in some parts of the parameter space.

The following condition quantifies the "strength" of identification of the entire identified set and modifies the standard condition for consistency in the point identifies case (see e.g. van der Vaart (1998), Theorem 5.52) or condition C.2 in Chernozhukov, Hong, and Tamer (2007) for the set-identified case.

Condition 4 (Polynomial Minorant) *There exist positive constants (κ_1, κ_2, r) such that for every $\varepsilon > 0$, there exists $\kappa_\varepsilon > 0$ such that for n large enough,*

$$\inf_{\theta\in\Theta:\varrho_n(\{\theta\},\Theta_{I,n})\geq\left(\frac{\kappa_\varepsilon}{n}\right)^r} \frac{\mu_n^{-1} n\hat{Q}_n(\theta)}{\left(\mu_n^{-1/r}\varrho_n(\{\theta\},\Theta_{I,n})\wedge\kappa_2\right)^r} \geq \kappa_1$$

with probability greater than $1-\varepsilon$.

Informally, we can read Condition 4 as putting a lower bound on the subgradients of the suitably normalized population criterion function over all points on the boundary of the identified set.[13] This is a direct analogue of the rank condition for identification in the point-identified case, as e.g. in Assumption 1 of Newey and Windmeijer (2008). Essentially this condition requires the rescaled signal part $\gamma_n(\theta)$ of the criterion to be bounded from below by a polynomial in the Euclidean distance of θ from the identification region $\Theta_{I,n}$.

[12] Note that if we let $\gamma_n(\theta) = Q_n(\theta)$ as defined in (1.2) and $\gamma(\theta) = Q_0(\theta)$ for the moment inequality setting in section 2, then by Theorem 7.31(b) in Rockafellar and Wets (1998) Condition 3(b) taken together with Condition 4 below ensures that for the ε_n blow-up of $\Theta_{I,n}$, $\limsup_n \Theta_{I,n}^{\varepsilon_n} \subset \Theta_I$ for any strictly positive sequence $\varepsilon_n \to 0$. On the other hand, by part (c) of the same Theorem, there exists a sequence $\varepsilon_n' \to 0$ such that $\limsup_n \Theta_{I,n}^{\varepsilon_n'} = \Theta_I$, so that Condition 1(b) holds. If $\Theta_{I,n}$ is nonincreasing in n, the second statement is clearly true for any positive null sequence ε_n, in which case Condition 1(b) will be redundant for the consistency result below.

[13] Since the population criterion function $\gamma_n(\theta)$ is typically not smooth on the boundary of the identified set, the gradient is not defined, so instead we have to consider the subgradient set $\partial Q(\theta)$, which is typically a convex cone, see Rockafellar and Wets (1998).

In order to analyze the convergence rate of the set estimator, we can now define the rate of "global" strength of identification

$$\mu_n := \lim_{\varepsilon\downarrow 0} \inf_{\theta\in\Theta_{I,n}^{\varepsilon}} \min_{j\leq k} \mu_{jn}(\theta) \tag{1.9}$$

where $A^{\varepsilon} := \{\theta \in \Theta : d(\theta, A) \leq \varepsilon\}$ denotes the closed ε-blowup of a set A. In the case of the linear IV model with a scalar endogenous regressor, μ_n corresponds to the rate of the concentration parameter. Note also that I allow the strength of identification to vary across the boundary of the identification region.[14]

To fix ideas, consider the most important special cases of this setup

1. the "classical" case of strong identification, which corresponds to $\mu_n = n$ and $m_n = m$. In the case of identification regions with a non-degenerate interior, Chernozhukov, Hong, and Tamer (2007) show that the set estimator defined below is $\sqrt{n}$-consistent with respect to the Hausdorff distance.

2. the set-identified version of weak identification with a fixed number of moment conditions, which is given by $m_n = m$ constant, and constant strength of mo-

[14] In order to see how this can happen in realistic applications, consider the following stylized example in the spirit of Manski and Pepper (2000)'s Monotone IV assumption:

Example 6 () (Bounds on the ATE in the Presence of Attrition) *Suppose we want to evaluate the effect of a binary treatment, $T_i \in \{0,1\}$ on a random variable with potential outcomes $Y_{it} = \alpha_i + \beta_i t$ under treatment $t = 0$ and 1, respectively. Suppose now that we have three different assignment mechanisms: $Z_i = 1$ corresponds to voluntary participation, $Z_i = 2$ to full compliance, and under $Z_i = 0$, all subjects are precluded from taking up the treatment, where we assume that the usual monotonicity condition holds, i.e. $P(D_{i0} \leq D_{i1} \leq D_{i2}) = 1$, where D_{ik} denotes the counterfactual treatment status under the treatment regime $Z_i = z_k$. To make the problem interesting, assume that there is also a problem with differential attrition, or some other violation of the exclusion restriction, such that $\mathbb{E}[Y_{it}|Z_{ik} = z]$ is increasing in $z \in \{0,1,2\}$ for $t = 0,1$. The effect of treatment on the outcome for individual i is given by $\beta_i = Y_{i1}^* - Y_{i0}^*$, and say we are interested in estimating the average treatment effect (ATE) for the non-attriting population under $Z_i = 1$ given by $\beta_0 := \mathbb{E}[\beta_i|Z_i = 1]$.*
Assuming that the average effect on the treated under voluntary participation is greater than the ATE (this could be justified e.g. by a Roy selection model), the moment restrictions implied by the model are $\mathbb{E}[(Y_i - \alpha)\mathbb{1}\{Z_i = 0\}] \leq 0$, $\mathbb{E}[(Y_i - \alpha - T_i\beta)\mathbb{1}\{Z_i = 1\}] \leq 0$, and $\mathbb{E}[(Y_i - \alpha - \beta)\mathbb{1}\{Z_i = 2\}] \geq 0$ where we can use the sample analogs to estimate the bounds.
Now, if under the voluntary treatment regime, take-up is very low, the upper bound on the ATE is only identified off a rather small group of "compliers" vis-à-vis the regime under which no subject receives treatment. On the other hand, the complier group corresponding to a change from voluntary participation to full compliance is then relatively large, so that identification of the upper bound is much weaker than that of the lower bound.

ments, $\mu_n = \mu$. Our results will show that for this case, the rescaled criterion has a non-deterministic limit, and the set estimator is inconsistent for any choice of critical values.

3. the many weak moments scenario corresponds to $\mu_n \to \infty$ and $m_n \to$. I will establish that if we have in addition that $\frac{m_n}{\mu_n} \to 0$, there is a consistent set estimator.

For a critical value c, we can define a set estimator as

$$\hat{\mathcal{C}}_n(c) = \{\theta \in \Theta : n\hat{Q}_n(\theta) \leq c\}$$

In order to ensure consistency, the critical value c_n should increase in sample size, and has to be chosen in a way such that $\hat{\mathcal{C}}_n(c)$ covers the identified set $\hat{\Theta}_{I,n}$ with probability approaching 1.

Condition 5 (Cutoff Value) *There is a sequence $\hat{c}_n$, which may depend on the data, such that (i) $\frac{\hat{c}_n}{\mu_n} \xrightarrow{p} 0$ and (ii) $P\left(\sup_{\theta\in\Theta_{I,n}} nQ_n(\theta) > \hat{c}_n\right) \to 0$*

The first part of Condition 5 requires the cut-off value to grow at a smaller rate than the rate of the signal component of the criterion function which, in conjunction with Condition 4, will force the set estimator to shrink towards the identified set from the outside. On the other hand, the second part of Condition 5 implies that the cut-off has to grow sufficiently fast to dominate the noise component in large samples. In general, there is no guarantee that such a sequence $\hat{c}_n$ exists, but I am going to give primitive sufficient conditions below in this section.

We can now state the general consistency result for the set estimator $\hat{\mathcal{C}}_n$:

Theorem 1 *(i) Suppose $\alpha_n \to 0$ and Conditions 1, 3, and 5 hold. Then $d_H(\hat{\mathcal{C}}_n, \Theta_{I,n}) \xrightarrow{p} 0$ so that $\hat{\mathcal{C}}_n$ is consistent. (ii) Suppose Conditions 1, 4, and 5 hold. Then $\hat{\mathcal{C}}_n$ is a consistent estimator for $\Theta_{I,n}$, and*

$$\hat{c}_n^{-1/r}\varrho_n(\hat{\mathcal{C}}_n(\hat{c}_n), \Theta_{I,n}) = O_P(1)$$

From Condition 2 and the second part of Theorem 1 we can now give the convergence rate of the set estimator with respect to the limiting identified set Θ_I.

Corollary 1 *Suppose Conditions 1, 2, 4, and 5 hold. Then*

$$\mu_n^{-1/r}\varrho_n(\hat{C}_n(\hat{c}_n), \Theta_I) = O_P\left(\left(\frac{\hat{c}_n}{\mu_n}\right)^{1/r} \vee \tau_n\right)$$

Example 7 (Conditional Moment Restrictions, continued) *Under the choice of basis functions discussed in Example 5, and noting that in this example $\mu_n = a_n^2$ and $r = 2$, the set estimator under a conditional moment restriction satisfies*

$$\mu_n^{-1/2}\varrho_n(\hat{C}_n(\hat{c}_n), \Theta_I) = O_P\left(\frac{\hat{c}_n \vee nm_n^{-4}}{\mu_n}\right)^{1/2}$$

If $m_n \to \infty$ and $\frac{m_n}{\mu_n} \to 0$, we can find a critical value for which $\frac{\hat{c}_n}{m_n} \to \infty$ satisfies Condition 5 and obtain a consistent estimator for the sharp identification region. In close analogy to more familiar problems in nonparametric estimation, we can interpret m_n^{-2} as the rate of the approximation "bias", and $\frac{\hat{c}_n}{n}$ as the rate of the "variance" contribution to the squared Hausdorff distance between the set estimator and the identification region, where relative to the standard setting, both parts are inflated by the factor $\frac{n}{\mu_n^2}$ accounting for "weaker than strong" identification.

If the distribution of the criterion does not degenerate in the interior of the sharp identification region, we can only bound the optimal rate for m_n^* by $m_n^* = o\left(n^{1/5} \vee \mu_n\right)$ since $\frac{\hat{c}_n}{m_n} \to 0$ (note also that in this example, the sequence μ_n doesn't depend on the number of moments). This bound depends on the rate of $\hat{c}_n$, and following Chernozhukov, Hong, and Tamer (2007) a feasible choice would be $\hat{c}_n = m_n \log n$, implying that $m_n^* = O\left(\frac{n}{\log n}\right)^{1/5}$. For strong identification, i.e. $\mu_n = n$, we can therefore bound the rate at which the set estimator converges in Hausdorff distance by $d_H(\hat{C}_n, \Theta_I) \geq O\left(n^{-2/5}\right)$. If the dimension of z is greater than 1, the "bias" term will vanish at a slower rate, so that the optimal number of moments will typically be greater than in the scalar case.

The previous example illustrates that in realistic cases, the information about the parameter (in this example the rate of the approximation error) from additional constraints can be quite small compared to their "cost" from adding noise to the estimation problem, so that keeping the number of moments small in small samples may in fact result in smaller set estimates or confidence regions.

1.3.1 Moment Inequality Model

I will now give primitive assumptions for the moment inequality model that are sufficient for the conditions for consistency of $\hat{\mathcal{C}}_n$. To fix notation, following Han and Phillips (2006), I will write the moment functions as the sum

$$g_n(Y_i, \theta) = \bar{g}_n(\theta) + \xi_n(Y_i, \theta)$$

where $\bar{g}_{mn}(\theta) = \mathbb{E}_{P_0}[g_{mn}(Y_i, \theta)]$ is the population expectation, and $\xi_{mn}(Y_i, \theta) = \hat{g}_{mn}(Y_i, \theta) - \bar{g}_{mn}(\theta))$ the noise component of the mth component of the moment vector for sample size n. Also define

$$\zeta_{mn}(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_{mn}(Y_i, \theta)$$

The partial derivatives of the moment functions $G_{ij}(\theta) = \frac{\partial}{\partial \theta_j} g(Y_i, \theta)$ are stacked into the matrix $G_i(\theta) = [G_{i1}(\theta), \dots, G_{ik}(\theta)]$. The average Jacobian is given by $\hat{G}(\theta) = \frac{1}{n} \sum_{i=1}^{n} G_i(\theta)$, and we denote the expected Jacobian by $\bar{G}(\theta) := \mathbb{E}_{P_0}[G_i(\theta)]$.

Assumption 1 (Set Identification) *(a) There are constants $\delta, C > 0$ such that for n large enough*

$$n \|\bar{g}_n(\theta)\|_{W,-} \geq C(\varrho_n(\theta, \Theta_{I,n}) \wedge \delta)^r$$

for all $\theta \in \Theta$, where $\|x\|_{W,-}$ denotes the Euclidean norm of the component-wise negative parts of a vector x given a weighting matrix W, and $\varrho_n(\cdot, \cdot)$ is as defined *in 1.8. (b) There is a sequence of constants $\mu_n \to \infty$ which is defined as*

$$\mu_n := \lim_{\varepsilon \downarrow 0} \inf_{\theta \in \Theta_{I,n}^{\varepsilon}} \min_{j \leq k} \mu_{jn}(\theta)$$

Note that if $\bar{g}_n(\theta)$ has uniformly continuous Jacobians $\bar{G}_n(\theta)$, Assumption 1 holds if the smallest eigenvalue of $H(\theta) = \lim_n nS_{n,\theta}^{-1}\bar{G}_n(\theta)\bar{G}(\theta)'S_{n,\theta}^{-1}$ is bounded away from zero uniformly over $\partial\Theta_{I,n}$ and $\min_{j \leq m_n} \inf_{\theta \in \partial\Theta_{I,n}} \mu_{jn}(\theta) \to \infty$. Note that for the point-identified case this corresponds to Assumption 1 in Newey and Windmeijer (2008).

We now state the main regularity assumptions on the signal component of the moment functions:

Assumption 2 (Moment Signal)

(a) The expectation of the moment functions $\bar{g}_n(\theta) = \mathbb{E}_{P_0}[g_{in}(\theta)]$ is continuous in $\theta \in \Theta$ for all n.

(b) The population criterion function

$$\gamma_n(\theta) := \frac{n}{\mu_n} Q_n(\theta) = \frac{n}{\mu_n} \min_{t \geq 0} (\bar{g}_n(\theta) - t)' W(\theta)(\bar{g}_n(\theta) - t)$$

is nonnegative and lower semi-continuous,

If the weighting matrix is diagonal, the "signal" $\gamma_n(\theta)$ from the moment restrictions is a weighted sum of the squared negative parts of the moment vector at θ. Note also that part (b) of Assumption 2 does not require $\gamma_n(\theta)$ to be finite in the limit. This is particularly important in the case in which the strength of identification varies across dimensions of the parameter space and different regions of the boundary of the identification region. We now state our main conditions on the noise component of the moment vector:

Assumption 3 (Moment Noise)

(a) For the rate of the number of moments, m_n, we have

$$\delta_n(\theta) = m_n^{-1}(n\hat{Q}_n(\theta) - \mu_n\gamma_n(\theta)) = O_p(1)$$

(b) $\delta_n(\theta) \xrightarrow{d} \delta(\theta)$ uniformly in θ.

(c) The first four moments of $\zeta_{mn}(\theta)$ are bounded uniformly in θ.

(d) $\max_{m\leq m_n} |\xi_{imn}(\theta)|$ is tight.

(e) The distribution of $\sup_{\theta\in\Theta_{I,n}} n\hat{Q}_n(\theta)$ is continuous.

All parts of Assumption 3 are fairly standard. I also impose a high-level assumption on the convergence of the weighting matrix in order to include the practically relevant case of a data-dependent choice for $W_n(\theta)$:

Assumption 4 *The weighting matrix $W_n(\theta)$ converges in probability to $W(\theta)$ in the sense that*

$$\max_{l,m\leq m_n} |w_{lm,n}(\theta) - w_{lm}(\theta)| \xrightarrow{p} 0$$

uniformly in θ, where $w_{lm,n}(\theta)$ and $w_{lm}(\theta)$ are the (l,m) elements of $W_n(\theta)$ and $W(\theta)$, respectively.

In most standard settings, a necessary condition for 4 to hold is that $\frac{m_n^2}{n} \to 0$. If $W_n(\theta)$ is the inverse of the variance-covariance matrix of the moment functions, we would have to require in addition that the fourth moments of $\xi_{in}(\theta)$ are bounded uniformly in $\theta \in \Theta$.[15]

I will now give the main condition on the relative rates of number and strength of moments:

Assumption 5 *$\mu_n \to \infty$ and $\alpha_n := \frac{m_n}{\mu_n} \to 0$ as $n \to \infty$.*

[15] Typically, in a setting with many moment conditions we would also care about higher-order efficiency of the estimated inverse variance matrix, as delivered by Empirical Likelihood (see e.g. Newey and Smith (2004)). However, in estimation using moment *inequalities*, the bias from estimating the slackness parameters is of the same order as that from estimating the Jacobian and the weighting matrix. Since no GEL criterion function appears to address the former problem, efficient weighting does not lead to an improvement in the rates for the set estimator.

For the classical linear instrumental variables problem Chao and Swanson (2005) showed that 2SLS is consistent under the rate satisfying Assumption 5, whereas LIML is consistent as long as $\frac{m_n^2}{\mu_n} \to 0$. As we will see below, in set-identified settings, inverse variance weighting will typically not achieve this improvement in rates because the "noise" component of the criterion will depend on the parameter θ not only through the variance of $\xi_{in}(\theta)$, but also through the slackness $[\bar{g}_n(\theta)]_+$ of the moment restrictions.

1.3.2 Criterion and Decomposition

We will now analyze consistency for the set estimator based on the Wald statistic

$$n\hat{Q}_n(\theta) := n \min_{t\geq 0} \|\hat{g}_n(\theta) - t\|^2_{W_n(\theta)} = \min_{t\geq 0} \|\sqrt{n}(\bar{g}_n(\theta) - t) + \zeta_n(\theta)\|^2_{W_n(\theta)}$$

for the moment inequality problem. Denoting the projection of a vector x onto a convex cone $\mathcal{C}$ with $\Pi(x|\mathcal{C}, W) := \arg\min_{t\in\mathcal{C}} \|x - t\|_W$, we define $\sqrt{n}t_{0n}(\theta) := \Pi(\sqrt{n}\bar{g}_n(\theta)|\mathbb{R}_+^{m_n}, W_{n,\theta})$ and $\sqrt{n}\hat{t}_n(\theta) := \Pi(\sqrt{n}\bar{g}_n(\theta) + \zeta_n(\theta)|\mathbb{R}_+^{m_n}, W_{n,\theta})$. Note that the projection $\Pi(x|\mathcal{C}, W)$ is well-defined and unique (see e.g. Theorem 1 in section 3.12 of Luenberger (1969)).

The following proposition states that under the assumptions made above, after proper rescaling, the criterion function converges uniformly to the decomposition into a signal and a deterministic noise component:

Proposition 1 *Under Assumptions 2-4, and $\alpha_n = \frac{m_n}{\mu_n} \to \alpha < \infty$,*

$$\sup_{\theta\in\Theta}(\mu_{n,\theta}h_n(\theta))^{-1}\left|n\hat{Q}_n(\theta) - \mu_{n,\theta}\left(\gamma_n(\theta) + \alpha_n\delta_n(\theta)\right)\right| \xrightarrow{p} 0$$

where $h_n(\theta) = 1 \vee \gamma_n(\theta)$, and

$$\delta_n(\theta) \;=\; \mathbb{E}\|\zeta_n(\theta) - \sqrt{n}(\hat{t}_n(\theta) - t_{0n}(\theta))\|^2_{W_{n,\theta}} - 2n(\bar{g}_n(\theta) - t_{0n}(\theta))'W_{n,\theta}(\mathbb{E}[\hat{t}_n(\theta)] - t_{0n}(\theta))$$

The first term of the noise component $\delta_n(\theta)$ is the expectation of a quadratic form of the projection residuals, which in the case of moment equalities[16] collapses to $\mathrm{tr}(W_{n,\theta}\Omega_n(\theta))$, the bias term of the standard GMM objective function (see e.g. Donald and Newey (2000)). Since a given moment only contributes to this bias term when it is binding, the bias on the criterion function is in a loose sense less severe than in the case of moment equalities. This discussion also suggests that we should expect finite sample bias to be more of a problem if the identification region is small or identification is weak in the sense that at all points of the identification region a large number of moment restrictions is close to binding. Note also that in the identified set, $\bar{g}_n(\theta) - t_{0n}(\theta) = 0$, so that the second term in δ_n is nonzero only outside of the identified set $\Theta_{I,n}$.

Lemma 1 *Suppose Assumptions 1-3 hold, and that there is a (possibly random) sequence $\hat{c}_n$ such that $\frac{\hat{c}_n}{\mu_n} \xrightarrow{p} 0$ and $\frac{m_n}{\hat{c}_n} \xrightarrow{p} 0$. Then $\hat{c}_n$ satisfies condition 5, and we have*

$$P\left(\sup_{\theta\in\Theta_{I,n}} n\hat{Q}_n(\theta) > \hat{c}_n \right) \to 0$$

Proof: By Assumption 1 (a), $\gamma_n(\theta) = 0$ for $\theta \in \Theta_{I,n}$. Hence, by Proposition 1,

$$\sup_{\theta\in\Theta_{I,n}} \mu_n^{-1}|n\hat{Q}_n(\theta) - m_n\delta_n(\theta)| \xrightarrow{p} 0, \text{ and } \sup_{\theta\in\Theta_{I,1}} |\delta_n(\theta)| \leq B$$

for some $B < \infty$ by Assumption 3. Therefore, for any $\eta, \varepsilon > 0$ and n large enough,

$$\begin{aligned} P(\sup_{\theta\in\Theta_{I,n}} \mu_n^{-1} n\hat{Q}_n(\theta) - \eta \leq \mu_n^{-1}\hat{c}_n) &\geq P(\sup_{\theta\in\Theta_{I,n}} \mu_n^{-1} m_n\delta_n(\theta) \leq \mu_n^{-1}\hat{c}_n) \\ &\geq P(m_n B \leq \hat{c}_n) > 1 - \varepsilon \end{aligned}$$

where the last step follows from $\frac{m_n}{\hat{c}_n} \xrightarrow{p} 0$, and $\sup_{\theta\in\Theta_{I,n}} \delta_n(\theta) \leq \sup_{\theta\in\Theta_{I,1}} \delta_n(\theta) < B$ where we used $\Theta_{I,n} \subset \Theta_{I,n}$ from Assumption 1. Since the choice of $\eta > 0$ was arbitrary, the result follows from Assumption 3 (b) □

[16] recall that we can represent any equality as a combination of two deterministically related inequalities

Since Assumption 5 ensures that a critical value $\hat{c}_n$ satisfying the assumptions of Lemma 1 exists, we can now state our main consistency result for the moment inequality model:

Theorem 2 *The Moment Inequality model given in Assumptions 1-5 satisfies Conditions 3-5. Hence Theorem 1 applies, and the set estimator $\hat{\mathcal{C}}_n$ is consistent.*

This result can be modified to accommodate moment selection procedures as in Andrews and Soares (2007), which can in many cases mitigate, but not entire solve, the problems with bias under many moment asymptotics. Also, while continuously updated inverse variance weighting is known to remove parts of the higher-order bias in GMM (see e.g. Chao and Swanson (2005) and Newey and Smith (2004)), for set estimation there will typically not be an improvement in the rate results as illustrated in the following example.

Example 8 (Linear "One-Sided" Instrumental Variables, continued) *For simplicity, assume that errors are independent of Z_i with $\mathrm{Var}(Y_i - X_i\theta|Z_i) = \sigma^2(\theta)$. Then it can be seen that for a weighting matrix of the form $W_n(\theta) = s_n(\theta)^2 \left(\frac{1}{n}Z'Z\right)^{-1}$, the noise component converges to $\delta(\theta) = \frac{\sigma^2(\theta)}{s_n(\theta)^2}H(\theta)$ for some function $H(\theta)$ which by inspection is minimized at some point in the identification region Θ_I (in the case of classical linear IV, $H(\theta) = 1$). Note that the latter depends crucially on the variance of the moment functions being a scalar multiple of $\frac{1}{n}Z'Z$ at any value of θ. By definition, $\sigma(\theta)^2$ is minimized at the probability limit of the OLS estimator, so that for $s_n(\theta) = \bar{s}$, a constant, $\mu_n^{-1}Q(\theta) = \gamma(\theta) + \alpha\frac{\sigma(\theta)^2}{\bar{s}^2}H(\theta)$ is minimized at a point which is "biased towards OLS" unless $\alpha = 0$. On the other hand, for continuously updated inverse variance weighting, $s_n(\theta)^2 = \sigma(\theta)^2$, the limiting criterion is minimized at some point in the identified set. However, in contrast to the point-identified case, this feature does not lead to an improvement in the fastest permissible rate for m_n as in Chao and Swanson (2005), but only guarantees that the limit of the set estimator has a nonempty intersection with the sharp identification region.*

1.4 Confidence Regions

In this section, I will show uniform validity of inference procedures using critical values obtained from "plug-in asymptotics" (henceforth PA) and subsampling. More specifically, we will consider the asymptotic confidence size of a nominal $1-\alpha$ confidence set $\hat{C}_n := \{\theta \in \Theta : \hat{T}_n(\theta) \leq c(\theta, 1-\alpha)\}$ based on a test statistic $\hat{T}_n(\theta)$ given a (possibly parameter-dependent) critical value $c(\theta, 1-\alpha)$.

Following Andrews and Guggenberger (2007b), we define the asymptotic confidence size of $\hat{C}_n$ as

$$AsyCS := \liminf_n \inf_{(\theta,P)\in\mathcal{P}_0} P\left(\hat{T}_n(\theta) \leq c(\theta, 1-\alpha)\right)$$

where $\mathcal{P}_0$ is the set of null distributions (θ, P) for $\theta \in \Theta_I(P)$, the identification region corresponding to the measure P.

By taking the infimum over (θ, P) before taking limits, this definition requires in particular that the underlying hypothesis test has size less than or equal to α uniformly in both the parameter of interest θ and other nuisance parameters of the distribution of the data, P. Uniformity with respect to θ is a minimal requirement for the correct coverage probabilities for the resulting confidence sets, and uniformity with respect to other features of P gives the procedure certain robustness properties, including robustness when identification is weak in the sense of the preceding discussion, or when the identification region is small, as discussed by Imbens and Manski (2004) and Stoye (2009).

Uniform validity of Gaussian asymptotic and subsampling procedures for inference with a finite number of moment conditions has been shown by Andrews and Guggenberger (2007b), and I am going to show how to modify and extend their arguments to situations with a growing number of moment inequalities.

1.4.1 Test Statistics

I now give a general framework of test functions $S(g, W)$ which depend on the (infinite-dimensional) moment vector g and a weighting operator W. The test function may depend on g or a suitable nondecreasing transformation $\varphi(g, m, n)$ of g which may vary with sample size n and the number of elements of g used for inference. This will make it possible to introduce a proper normalization of the moment functions as well as incorporate generalized moment selection procedures as in Andrews and Soares (2007) into our framework. In order to account for the fact that only m moment conditions are used for inference, I will consider the component-wise transformation $\varphi_{mn}(g)$ whose lth element is given by

$$\varphi^{(1)}_{nm,l}(g) := g_l \mathbb{1}\{l \leq m\}$$

where m is the number of moments used for inference.

The weighting operator $W : l_2 \times l_2 \to \mathbb{R}$ is a positive definite bilinear mapping on the space of square-summable sequences in $\mathbb{R}$ (a bilinear map is said to be positive definite if for any $x \in l_2$, $W(x, x) \geq 0$, where the inequality is strict if $x \neq 0$). In the Hilbert space l_2 endowed with the norm induced by the usual scalar product,

$$W(x, y) = \langle x, y \rangle_W = \langle x, Wy \rangle = \sum_{i,j \geq 1} x_i w_{ij} y_j \tag{1.10}$$

so that the weighting function can be represented by the linear operator W. In order to operationalize convergence of bilinear forms, we will use the metric induced by the operator norm for W in l_2,

$$d(W_1, W_2) := \sup_{\|x\| \leq 1} \|(W_1 - W_2)x\|$$

The weighting operator W is a member of $\Psi \subset \{B : l_2 \to \mathbb{R}$ such that $q(x) = \langle x, Wx \rangle$ positive definite$\} \subset B(l_2)$, the space of bounded, self-adjoint linear operators on l_2.[17]

Since in finite samples only a finite-dimensional subvector of g is used for inference, in some cases, the weighting matrix is replaced by $\psi_1(W_n, m, n) := \left(\left\{w^{kl}\mathbb{1}\{k, l \leq m\}\right\}_{k,l\geq 1}\right)^{-1}$, where w^{kl} is the (k,l)th element of the inverse of W. This mainly concerns the case $W_n(\theta) = \Omega(\theta)^{-1}$, and since this transformation preserves continuity in nuisance parameters and positive semi-definiteness, I will suppress the function $\psi(\cdot)$ in the subsequent discussion.

Given a choice of a test function $S(\cdot,\cdot)$, we will consider inference based on the statistic

$$\hat{T}_{nm}(\theta) = a_m S(\varphi^{(1)}_{nm}(\hat{g}_n(\theta)), W_n(\theta))$$

where m_n is the number of moment conditions used for inference, and a_m is a sequence of known normalizing constants which will ensure that the distribution of the test statistic does not degenerate as the number of moments grows. Note that the mean of the distribution of $\hat{T}_n(\theta)$ will typically depend on the slackness of the constraints in a complicated manner, and may well diverge as the number of moments grows. However, this turns out not to be relevant for the uniform coverage results presented in this section, and I will therefore address this point only for the distribution under the least favorable hypothesis.

1.4.2 Examples for Test Functions

A commonly used test function penalizes the one-sided deviations of the sample moment functions (see e.g. Manski and Tamer (2002)) and can be extended to

$$S_1(g, W) = \sum_{l\geq 1}\left[\frac{g_l}{\hat{\sigma}_{ll}}\right]_-^2$$

[17] Note that below, W will only operate on differences $(g-t)$ for some nonnegative sequence t. Even though the moment vector g need not be square-summable, for the value of t solving the optimization problem implicit in the computation of each of the statistics below, we will have $(g-t) \in l_2$ with probability 1 under the null hypothesis.

where $\sigma_{ll}(\theta) = \sqrt{\mathrm{Var}(g_{il}(\theta))}$ so that the corresponding test statistic takes the form

$$\hat{T}_{nm,1}(\theta) = a_m S_1(\varphi_{nm}^{(1)}(\hat{g}_n(\theta)), W_n(\theta)) = a_m n \sum_{l=1}^{m} \left| \frac{\hat{g}_{ln}(\theta)}{\hat{\sigma}_{ll,n}(\theta)} \right|_-^2$$

For a fixed value of m, this statistic coincides with that defined by the function $S_1(\cdot,\cdot)$ in Andrews and Guggenberger (2007b).

The second statistic of interest is an extension of the quasi-likelihood ratio statistic (QLR, see Silvapulle and Sen (2005)), which has also been applied to the problem of set inference based on moment inequalities, see Rosen (2008). I consider a modification which allows for a sequence of moment functions,

$$S_2(g, W) = \min_{t \geq 0} \|g - t\|_W^2$$

so that the corresponding test statistic takes the form

$$\hat{T}_{nm,2}(\theta) = a_m S_2(\varphi_{nm}^{(1)}(\hat{g}_n(\theta)), W_n(\theta)) = a_m n \min_{t \geq 0} (\hat{g}_{m,n}(\theta) - t)' W_{m,n}(\theta)(\hat{g}_{m,n}(\theta) - t)$$

where $\hat{g}_{m,n}$ denotes the subvector consisting of the first m components of $\hat{g}_n$, and $W_{m,n}$ denotes the corresponding $m \times m$ submatrix of W_n.[18]

Note that in principle, evaluating the test function S_2 involves a minimization over an infinite-dimensional parameter (see e.g. Luenberger (1969) or Rockafellar and Wets (1998)), but in finite samples we will only have to deal with the finite-dimensional version of this problem, since for each n we use only a finite number of moments for inference. Either statistic can be combined with a moment selection procedure like the one suggested by Andrews and Soares (2007) to improve power in cases for which some moment constraints are very slack for some parameter values θ.

Furthermore, I consider the Generalized Empirical Likelihood Ratio (GELR) statistic which for point-identified problems defines the class of GEL estimators analyzed

[18] Variations of this statistic with different choices for the weighting matrix and the cone for t have been analyzed frequently in the literature, e.g. if we replace the maximization over $t \in \mathbb{R}_+^\infty$ with $t \in \mathcal{C}_W := W(\theta)^{-1/2}\mathbb{R}_+^\infty$, we obtain the weighted GMM statistic considered in e.g. Pakes, Porter, Ho, and Ishii (2006) or Chernozhukov, Hong, and Tamer (2007). $S_2(\hat{g}_n(\theta), \Omega_n(\theta)^{-1})$ is the QLR statistic.

by Newey and Smith (2004). One of the most prominent subcases is the Empirical Likelihood Ratio (ELR) statistic for which Canay (2007) showed large-deviation optimality for tests under moment inequalities. The GELR statistic is given by

$$\hat{T}_n^{GELR}(\theta) := \inf_{t\geq 0} \sup_{\lambda\in\hat{\Lambda}(\theta,t)} n\hat{P}(\lambda,\theta,t) = \inf_{t\geq 0} \sup_{\lambda\in\hat{\Lambda}(\theta,t)} \sum_{i=1}^{n} \varrho\left(\langle\lambda, g_{in}(\theta)-t\rangle\right) - n\varrho(0)$$

where $\varrho(v)$ is a strictly concave function of v, $\hat{P}(\lambda,\theta,t) = \frac{1}{n}\sum_{i=1}^{n}\varrho\left(\langle\lambda, g_{in}(\theta)-t\rangle\right) - \varrho(0)$, and $\hat{\Lambda}_n(\theta,t) = \{\lambda\in l_2 : \langle\lambda, g_{in}(\theta)-t\rangle \in \text{dom }\varrho\}$. For $\varrho(v) = \log(1+v)$, the GELR statistic corresponds to the Empirical Likelihood Ratio statistic, and if the data are i.i.d. and $\varrho(v) = -\frac{(1+v)^2}{2}$, this becomes a feasible QLR statistic with $W_n(\theta) = \hat{\Omega}_n(\theta)$, see e.g. Newey and Smith (2004). Even though the GELR statistic can't be expressed directly in terms of a test function of the average moment vector $\hat{g}_n$ and a weighting matrix, in section 4 we will give conditions under which the GELR statistic is asymptotically equivalent to $S_2(g,\Omega(\theta)^{-1})$ under many moments asymptotics.

Another test function of interest is

$$S_3(g,W) = \sup_{a\in A\subset\mathcal{C}_W} |\langle a, W^{1/2}g\rangle|_-^2$$

where we take the supremum over certain positive linear combinations of the moment functions. E.g. for $W = I$ and $A = \{e_1, e_2, \dots\}$, this statistic simply penalizes the largest violation in the set of constraints. The Kolmogorov-Smirnov type statistic for countable intersection bounds arising from the conditional moment inequality model in Chernozhukov, Lee, and Rosen (2008) using series approximations falls into this class.

1.4.3 Main Assumptions

I will now state basic conditions on the test functions used for the construction of confidence sets. Below, we will show that all test functions discussed in Section 2 satisfy these requirements.

Condition 6 *The statistic of interest can be expressed as* $\hat{T}_{nm}(\theta) = a_m(S(\tilde{g}_{nm}(\theta), W_n(\theta)) - B_m) + o_P(1)$*, where*

(a) The statistic $S(g, W)$ *is nonincreasing in* g*.*

(b) $S(g, W)$ *is continuous at* $g \in l_2$ *and* $W \in \Psi$*.*

(c) $S(\Delta g, \Delta^{-1} W \Delta^{-1}) = S(g, W)$ *for all* $g \in l_2$*,* $W \in \Psi$ *and p.d. diagonal* Δ*.*

(d) $S(g, W) \geq 0$ *for all* g *and* W *positive definite.*

(e) $S(g, W)$ *is quasi-convex in* g *for* W *positive definite.*

Note that for any continuous nondecreasing function $\varphi : l_2 \longrightarrow l_2$, $S(\varphi(g), W)$ inherits properties (a),(b),(d) and (e) from $S(g, W)$.

For the following condition, let $\varphi(g)$ be the subvector obtained from g by eliminating all components m with $h_{1m} = \infty$, and let $\psi(W)$ be either the sub-matrix of W corresponding to the elements in $\varphi(g)$, or the inverse of the corresponding sub-matrix of W^{-1} if $W = \Omega^{-1}$.

Condition 7 *For all positive sequences (with some elements potentially being infinite)* $h_1 \in \mathbb{R}^{\infty}_{+,\infty}$*, all* $W \in \Psi$*, and Gaussian sequence* Z *with mean zero and covariance operator* Ω*, the distribution function of* $S(Z + h_1, W)$ *at* $t \in \mathbb{R}$ *is*

(a) continuous for $t > 0$

(b) strictly increasing for $t > 0$ *unless* $h_{1m} = \infty$ *for all* $m = 1, 2, \ldots$

(c) less than or equal to $\frac{1}{2}$ *at* $t = 0$ *whenever* $h_{1m} = 0$ *for all* $m = 1, 2, \ldots$*.*

(d) For the selection functions $\varphi(\cdot)$ *and* $\psi(\cdot)$ *defined above,* $S(\varphi(g), \psi(W)) = S(g, W)$*.*

Note that part (a) and (b) require that the statistic is normalized properly thus ensuring that its distribution doesn't degenerate at any point on the positive real axis, except potentially at zero.

The moment functions will be required to have uniformly bounded fourth moments in order to allow for a Gaussian approximation and consistent estimation of the covariance matrix under the restrictions on the rate for the number of moments given below. The latter is necessary for the calculation of PA critical values, and in some cases the weighting operator may also depend on estimated components of the covariance operator.

Condition 8 *There exists a constant C such that for all m and n $\mathbb{E}|g_{imn}(\theta) - \bar{g}_{mn}(\theta)|^4 < C$ uniformly in θ.*

The rate at which we can allow the number of moments to grow has to be slow enough to ensure that we can approximate the distribution of the moment vector by a Gaussian is given by the following condition.

Condition 9 *The growth rate of the number of moments satisfies $m_n \longrightarrow \infty$ and $\frac{m_n^7}{n^2} \longrightarrow 0$.*

Note that this rate condition is much more restrictive than that needed for a normal approximation in the point-identified case, e.g. Newey and Windmeijer (2008) show that if the moment functions are uniformly bounded, the AR statistic can be approximated under the null hypothesis by a chi-squared with m_n degrees of freedom as long as $\frac{m_n}{n} \longrightarrow 0$. For moment inequalities, the chi-bar square approximation is a weighted average of chi-squared random variables with degrees of freedom less than or equal to m_n (see e.g. Silvapulle and Sen (2005)). However, this approximation relies heavily on the rotational symmetry of the distribution of Gaussian random vectors and therefore only valid if the moment vector is centered at the origin and not approximated well enough by a multivariate normal distribution. For the more general case, the distribution of the length of residuals from projections of random vectors onto convex cones is not well understood, which does not rule out that the rate stated in the previous condition may be improved upon for many instances of the test function $S(\cdot, \cdot)$.

Example 9 (Conditional Moment Restrictions, continued) *The rate condition need not be restrictive if we are mainly interested in inference on the sharp identification* *region. The optimal rate for consistency derived for set estimation under a conditional moment inequality in Example 7, $m_n = o\left(n^{1/5}\right)$, satisfies Condition 9. Hence, according to our main results on inference with moment inequalities below, the rate needed for the Gaussian approximation to be valid does not impose any additional restrictions if the number of unconditional moments is chosen as to ensure the fastest possible rate of convergence for the set estimator discussed in Section 3. However, if the conditioning variable has dimension greater than 1, the curse of dimensionality in the approximation error for the function $h(z, \theta, P)$ may make Condition 9 the binding constraint on the number of moments.*

For a given sample size n, we will parameterize the null distributions $(\theta, P) = (\theta, P_h) \in \mathcal{P}_0$ by θ and a vector $h \in H$ for some appropriately chosen index set H.[19] h can be split into three components h_1, h_2, and h_3, where h_1 contains the slackness parameters of the moment inequalities, $h_1 = n^{1/2}\sqrt{n}\mathbb{E}[g_{im}(\theta)]$, where the some constraints may be close to binding at θ, so that this limit may be finite. The vector h_2 contains auxiliary parameters that have to be estimated to obtain the weighting operator W, and we will assume throughout that all components of h_2 can be estimated consistently. h_3 captures other features of the underlying population distribution P and may be infinite-dimensional. This distinction is important when we analyze subsampling procedures since in many instances, the subsampling distribution gives a poor approximation to features of the population distribution which are best modeled by local parameters (see e.g. Mikusheva (2007) and Andrews and Guggenberger (2007a)).

We will now consider two different procedures to obtain critical values for the construction of confidence intervals:

[19] Note that $\mathcal{P}_0$ does not have a cartesian product form in the coordinate pairs (θ, P), since the set of possible values θ is given by the identification region $\Theta_I := \Theta_I(P)$ and therefore depends on the choice of P.

1.4.4 "Plug-In Asymptotic" Critical Values

The PA critical value $c_F(\theta, 1-\alpha)$ is computed using a consistent estimator $\hat{h}_2$ for the nuisance parameters h_2 and replacing the component of the nuisance parameter vector h_1 which cannot be estimated consistently with the values corresponding to the least favorable hypothesis, usually $h_1 = 0$. More specifically, let

$$T_n^*(h_1, h_2, \theta) := a_{m_n} S(\phi_{nm_n}(h_1 + Z_n), W_n) \tag{1.11}$$

where Z_n is a Gaussian vector with mean zero and covariance operator Ω_{n,h_2}. Since by condition 7 (a), the distribution of $T_n^*(h_1, h_2, \theta)$ is continuous for $t > 0$, we can choose the plug-in asymptotic critical value $\hat{c}_F(\theta, 1-\alpha)$ as the smallest value c such that

$$P(T_n^*(0, \hat{\eta}_{2n}, \theta) \leq c) \geq 1 - \alpha$$

In practice, one obtains $\hat{c}_F(\theta, 1-\alpha)$ as the $1-\alpha$ quantile of a simulated sample of $T_n^*(0, \hat{h}_2, \theta)$ based on Gaussian random draws Z_n.

In order for the procedure based on PA critical values to be similar on the boundary of the null hypothesis, we need the following condition to hold:

Condition 10 *For some $(\theta, P) \in \mathcal{P}_0$ with $h_1(\theta, P) = 0$, the distribution function of $S(Z, W(\theta, P))$ is continuous at its $1-\alpha$ quantile, where Z is a mean zero Gaussian sequence with covariance operator $\Omega(\theta, P)$.*

This condition requires that the postulated least favorable value of h_1 is in fact attained by at least one member in the family of probability measures $\mathcal{P}$.

1.4.5 Subsampling Critical Values

For block size $b < n$, we define the jth subsample statistic given the test function S as

$$\hat{T}_{nmbj}(\theta) = a_m \left(S(\varphi^{(1)}_{bm}(\hat{g}_{bj}(\theta)), W_n(\theta)) - B_m \right)$$

where the subscript j indicates that the moment function is evaluated at the jth subsample of size b. Note that since the normalizing constants (a_m, B_m) depend on sample size only through the number of moments, and are therefore the same as for the full-sample statistic $\hat{T}_{nm}(\theta)$.

The subsampling approximation to the distribution of $\hat{T}_{nm}(\theta)$ is then constructed using the c.d.f. for the subsample statistic $\hat{T}_{nmbj}(\theta)$ over the N_{nb} subsamples,

$$L_{nmb}(\theta, t) := N_{nb}^{-1} \sum_{j=1}^{N_{nb}} \mathbb{1}\left\{\hat{T}_{nmbj}(\theta) \leq t\right\}$$

where in the case of independent samples, $N_{nb} = \binom{n}{b}$ is the number of subsets of $W_1, \ldots, W_n$ of size b. The subsampling critical value $\hat{c}_S(\theta, 1-\alpha)$ at θ is the smallest value of c such that $L_{nmb}(\theta, c) \geq 1 - \alpha$.

The subsample size has to satisfy the following requirements:

Condition 11 *The subsample size b_n satisfies $b_n \to \infty$, $\frac{b_n}{n} \to 0$, and $\frac{m_n^7}{b_n^2} \to 0$.*

Note that for estimation of first and second moments of a finite-dimensional distribution, the optimal rate for b_n is typically of the order $n^{1/3}$ (see e.g. Politis, Romano, and Wolf (1999) and references therein), for which the rate of the number of moments would have to satisfy $\frac{m_n^{21}}{n^2} \to 0$, a third of the rate needed for inference based on plug-in asymptotics.

The following condition will be needed to establish that inference using subsampling critical values is non-conservative in the sense that for at least one distribution in $\mathcal{P}_0$, the asymptotic size of the subsampling confidence set is equal to its nominal level:

Condition 12 *For some $(\theta, P) \in \mathcal{P}$, the distribution function of $S(Z+h_1(\theta, P), W(\theta, P))$ is continuous at its $1-\alpha$ quantile, where Z is a mean zero Gaussian sequence with covariance operator $\Omega_{h_2(\theta,P)}$.*

1.4.6 Main Results

We will first state a preliminary result which will both be used for the proof of the main theorems, but also justifies the use of a particular instance of the generalized moment selection procedure proposed by Andrews and Soares (2007). Let ϱ_n be a sequence such that

$$\limsup_n \frac{\varrho_n}{(2\log\log n)^{1/2}} > 1$$

Define $\varphi(g)$ and $\psi(W)$ as in Condition 7 (d), and let $\varphi_n(g)$ be the vector obtained from g by deleting all components m such that $g_m > \varrho_n$, and $\psi_n(W)$ the components of W corresponding to the elements of $\varphi_n(g)$. Then we can state the following result:

Proposition 2 *Under Conditions 6, 7, and 8*

$$\limsup_n P(S(\varphi_n(\hat{g}_n), \psi_n(W_n)) > S(\hat{g}_n, W_n)) = \limsup_n P(S(\varphi_n(\hat{g}_n), \psi_n(W_n)) < S(\hat{g}_n, W_n)) = 0$$

Due to the "liminf" in the definition of asymptotic confidence size, it is not sufficient to consider pointwise limits at the parameter h of interest, but limits along subsequences w_n of sample size and all parameter sequences which converge to h. For this argument we will use the following notation from Andrews and Guggenberger (2007b):

Notation 1 *As in Andrews and Guggenberger (2007b), we define the sequences* $\eta_n = (\eta_{1n}, \eta_{2n}, \eta_{3n})$ *and* $\gamma_n = (\gamma_{1n}, \gamma_{2n}, \gamma_{3n})$, $(\gamma_n, \eta_n) \in \mathbb{R}^{\infty}_{\infty} \times \Psi$ *for all* n, *where for a given sequence* b_n, $\gamma_{2n} = \eta_{2n}$, $\eta_{1n} := n^{1/2}\bar{g}_n(\theta_n)$, *and* $\gamma_{1n} := b_n^{1/2}\bar{g}_n(\theta_n)$.

Convergence along all subsequences ensures that the limsup and the liminf of finite-sample confidence sizes coincide and determine the asymptotic confidence size as defined above. The following is the main coverage result for plug-in critical values under many moment asymptotics:

Theorem 3 *Suppose Conditions 6-9 hold. Then for* $0 < \alpha < \frac{1}{2}$, *the nominal level* $1 - \alpha$ *confidence set based on* $\hat{T}_n(\theta)$ *and critical values* $\hat{c}_n(\theta, 1 - \alpha)$ *obtained from* *plug-in asymptotics satisfies*

$$\liminf_n \inf_{(\theta,P)} P(\hat{T}_n(\theta) \leq \hat{c}_F(\theta, 1-\alpha)) \geq 1-\alpha$$

If in addition Condition 10 holds,

$$\liminf_n \inf_{(\theta,P)} P(\hat{T}_n(\theta) \leq \hat{c}_F(\theta, 1-\alpha)) = 1-\alpha$$

The argument behind this result is similar to that of Theorem 2 in Andrews and Guggenberger (2007b), however we have to account for the fact that under reasonable conditions, the distribution of the test statistic need not converge to a proper limit. We can also no longer rely on finite-dimensional convergence results for the moment functions, and we have to re-normalize the sequences and use a truncation argument in order to ensure that the statistic is properly defined for an increasing number of moments. Under regularity conditions, we can now give a uniform coverage result for subsampling critical values:

Theorem 4 *Suppose Conditions 6-9, and 11 hold. Then for $0 < \alpha < \frac{1}{2}$, the nominal level $1-\alpha$ confidence set based on $T_n(\theta)$ and critical values $\hat{c}_n(\theta, 1-\alpha)$ obtained from subsampling satisfies*

$$\liminf_n \inf_{(\theta,F)} P_F(\hat{T}_n(\theta) \leq \hat{c}_S(\theta, 1-\alpha)) \geq 1-\alpha$$

If in addition Condition 12 holds,

$$\liminf_n \inf_{(\theta,F)} P_F(\hat{T}_n(\theta) \leq \hat{c}_S(\theta, 1-\alpha)) = 1-\alpha$$

The rate condition on m_n needed for Theorem 4 is much stronger than that for Theorem 3. This is a direct result of the fact that the distribution of the statistic also depends on the nuisance parameters h_3 characterizing distributional features of the moment functions other than the first and second moments. Unlike in the case of finitely many moments, the corresponding parameters of the sample distribution do not necessarily converge to the values corresponding to a Gaussian limiting distribution, but the Gaussian approximation will generally only be valid under the rate restrictions on m_n relative to sample size n or subsample size b_n, respectively. If the corresponding components of the nuisance parameter vector converge to different limits for the sample and the subsampling distributions, there is no guarantee that the critical values obtained from subsampling will be conservative.

It should also be noted that the rate conditions for Theorem 3 and Theorem 4 are sufficient but not sharp. Below in Proposition 4 we will give sharp rates for two important subcases which considerably weaker than in the general case, but still restrict the growth rates in m_n relative to n and b_n, respectively. Hence, there is a range of growth rates in the number of moment conditions for which the Gaussian approximation works, but subsampling does not.

As a final remark on the general inference result, I should point out that the theoretical argument justifying the Generalized Moment Selection (GMS) procedure suggested by Andrews and Soares (2007) and Bugni (2008) can also be extended to many moment asymptotics along the lines of the previous argument.[20] This insight is of great practical importance because the power advantage of moment selection procedures should be expected to play out particulary strongly in testing problems involving a very large number of constraints.

1.4.7 Asymptotic Results for Quadratic Forms with Gaussian Errors

We will now state different sets of conditions under which the quadratic forms corresponding to the test functions S_1 and S_2 converge in distribution to a normal random variable under the least favorable hypothesis. The limiting distribution of the test statistics under any value of the nuisance parameter pertaining to the null hypothesis

[20] More specifically, under quasi-convexity of the test function and imposing the rate condition needed for the Berry-Esséen bounds, the proof of their Theorem 1 on asymptotically correct coverage of GMS confidence sets goes through using the same truncation and approximation arguments as in the proof of Theorem 3 in this paper. In order to avoid unnecessary additional notation, I will not reproduce the proof, but refer the reader to the proofs in Andrews and Soares (2007).

will therefore be dominated by a normal experiment.

We will first discuss the case in which the joint finite-sample distribution of the sample moment functions is Gaussian. Under the conditions for Theorem 3, the distribution of the statistic for non-Gaussian data will be approximated by that for a Gaussian experiment, so that the limiting argument will continue to hold.

For the asymptotic normality result, we will apply a central limit theorem for heterogeneous strong mixing sequences to the distribution of chi-bar square weights for the QLR statistic. Recall that the chi-bar square weight for j degrees of freedom is equal to the probability that exactly j constraints are binding (see Kudo (1963)). For the mth moment condition, define $D_{lm}(\theta) := \mathbb{1}\{Z_l - \Pi_l(Z|\mathbb{R}^m_+, \Omega_{nm}(\theta)) < 0\}$. Then the number of binding constraints is given by $\sum_{l=1}^m D_{lm}(\theta)$. If the moment functions are Gaussian, by a result from Kolmogorov and Rozanov (1960), we can give sufficient conditions for strong mixing of $D_{lm}(\theta)$ in terms of the second moments of $\zeta(\theta) := \hat{g}_n(\theta) - \bar{g}_n(\theta)$:

Condition 13 *(i) $\zeta = (\zeta_1, \dots, \zeta_m)$ is an m-dimensional random vector with $\mathbb{E}[\zeta_l] = 0$ for $l = 1, \dots, m$, $\mathbb{E}[\zeta\zeta'] = \Omega_{nm}(\theta)$, (ii) the eigenvalues of $\sup_m \text{eig}\Omega_{nm}(\theta) < B$ for some $B < \infty$, and (iii) for ω_{kl}, the (k,l)th element of Ω we have $\omega_{kl} = o\left(|k-l|^{-2}\right)$.*

Geometrically, this condition also implies that neither the cone $\mathcal{C}_{\Omega^{-1}}$ corresponding to the null hypothesis nor its polar cone become "too small" as we add more moment conditions, so the distribution of chi-bar square weights does not degenerate.

Now let

$$\bar{\sigma}_m(\theta)^2 := \text{Var}\left(m^{-1/2}\sum_{l=1}^m D_{lm}(\theta)\right) \tag{1.12}$$

Now we can show the following limiting result for the QLR statistic as the number m of moment restrictions goes to infinity:

Proposition 3 *Suppose Condition 13 holds and ζ is Gaussian, then the QLR statistic under the least favorable hypothesis, $\hat{T}_{nm}(\theta) = a_m \min_{t\geq 0}(\zeta_n(\theta)-t)'\Omega_{nm}(\theta)^{-1}(\zeta_n(\theta)-t)$ converges in distribution to*

$$\frac{S(\varphi_{nm}(\zeta_n(\theta)), \Omega_{nm}(\theta)^{-1}) - \frac{m}{2}}{\sqrt{m(1+\bar{\sigma}_m(\theta)^2)}} \xrightarrow{d} N(0,1)$$

Example 10 (Diagonal Covariance Matrix) *Suppose under the least favorable hypothesis $\hat{g}_n(\theta) \sim N(0, \Omega_n(\theta))$, where $\Omega_n(\theta) = \text{diag}(\omega_{1n}(\theta), \omega_{2n}(\theta), \dots)$ is diagonal. Then the finite-sample distribution of $\hat{T}_n(\theta)$ is chi-bar squared with c.d.f.*

$$P(\hat{T}_n(\theta) \geq t) = \sum_{j=1}^{m_n} w_j(\theta) P(\chi_j^2 \geq c) = \sum_{j=1}^{m_n} 2^{-m_n} \binom{m_n}{j} P(\chi_j^2 \geq c)$$

where χ_j^2 is a chi-squared random variable with j degrees of freedom. As shown in the appendix, for $m_n \to \infty$ we have

$$\frac{2\hat{T}_n(\theta) - m_n}{\sqrt{m_n}} \xrightarrow{d} N(0, 5)$$

1.4.8 Results for Commonly Used Statistics

We will now turn to the statistics given by the test functions $S_1 - S_3$ and the GELR statistic, and show how they fit into the general framework for which we derived the general inference results above.

Lemma 2 *(i) Under Condition 13, the statistics corresponding to the test functions S_1 and S_2 satisfy Conditions 6 and 7 with $a_m = \tilde{m}^{-1/2}$, where $\tilde{m}$ corresponds to the number of elements of the subvector $(h_{11}, \dots, h_{1m})'$ that are finite. (ii) The test function S_3 also satisfies Condition 6.*

It is important to point out the role of the correlation structure among the moments for the choice of a_m. The most commonly used forms of Cramér-van-Mises type statistics in the literature (see van der Vaart (1998) and also Domínguez and Lobato (2004) or Linton, Song, and Whang (2008) for examples) are based on the empirical c.d.f. and converge to functionals of a (non-ergodic) Brownian bridge. In those instances the proper normalizing constant is $a_m = \tilde{m}^{-1}$, whereas the rate $a_m = \tilde{m}^{-1/2}$ for the statistics defined by the quadratic forms S_1 or S_2 depends crucially on the ergodicity assumption in Condition 13.

We will now give conditions under which the GELR statistic satisfies Conditions 6 and 7:

Assumption 6 *(a) The variance of $\xi_{imn}(\theta)$ is bounded away from zero and from above uniformly for all m, n, and Condition 14 holds, and (b) $\hat{t}_n := \arg\inf_{t \geq 0} \|\hat{g}_n(\theta) - t\|^2_{\hat{\Omega}(\theta)^{-1}}$ and $\hat{t}_n^* := \arg\inf_{t\geq 0} \sup_{\lambda \in \hat{\Lambda}_n(\theta,t)} \hat{P}_n(\lambda, \theta, t)$ are defined for all distributions in $\mathcal{M}$, and $m_n^{-1}\|\hat{t}_n\|^2$ and $m_n^{-1}\|\hat{t}_n^*\|^2$ are uniformly bounded for all n with probability 1.*

Lemma 3 *Under Assumption 6, we can approximate*

$$\hat{T}_n^{GELR}(\theta) = S_2(\hat{g}_n(\theta), \Omega_n(\theta)^{-1}) + O_P\left(\frac{m_n}{n}\right)$$

In particular, if $\frac{m_n^2}{n} \to 0$, by Lemma 2 for S_2, the GELR statistic satisfies conditions 6 and 7.

The proof of this Lemma follows exactly the same logic as the argument in section 10.3 Andrews and Guggenberger (2007b) and will therefore be omitted. The only modifications needed for many moment asymptotics are that $\|\hat{g}_n(\theta) - \hat{t}_n\| = O_P\left(\frac{m_n}{n}\right)$, $\|\hat{t}_n - t_0\| = O_P\left(\frac{m_n}{n}\right)$, and that we need $\frac{m_n^2}{n} \to 0$ for consistency of the weighting matrix. This establishes $\hat{T}_n^{GELR}(\theta) - S_2(\hat{g}_n(\theta), \hat{\Omega}_n(\theta)^{-1}) = O_P\left(\frac{m_n}{n}\right) = o_P(1)$ by the assumptions of the Lemma. Hence we can apply Lemma 2 to establish Conditions 6-7 for the GEL statistic.

We now state conditions under which it is possible to implement feasible inverse variance weighting for statistics based on the test functions S_1 and S_2. This requires that the top left $m_n \times m_n$ submatrix of the covariance operator $\Omega_n(\theta)$ can be estimated consistently.

Condition 14 *(a) $n\{\mathbb{E}[\hat{g}_n(\theta)\hat{g}_n(\theta)'] - \mathbb{E}[\hat{g}_n(\theta)]\mathbb{E}[\hat{g}_n(\theta)']\} = \Omega(\theta)$*

(b) There exists a consistent estimator $\hat{\Omega}_{m,n}(\theta)$, i.e.

$$\hat{\Omega}_{m_n,n}(\theta) - \Omega_{m_n}(\theta) \xrightarrow{p} 0$$

(c) The eigenvalues of $\Omega_m(\theta)$ are bounded from below uniformly in m and θ.

(d) The elements of $\Omega_m(\theta)$ are bounded in absolute value uniformly in m and θ.

Under condition 14 and $\frac{m_n^2}{n} \to 0$, the weighting matrix $\psi(\Omega_n(\theta).m,n)$ can be estimated consistently, and using the results from Proposition 2 and Lemma 3, we can give a coverage result for some of the most instances of the general framework for inference set out above.

Corollary 2 *Suppose $W_n(\theta) = \Omega_n(\theta)^{-1}$. For the test statistics $\hat{T}_{n,1}(\theta)$ and $\hat{T}_{n,2}(\theta)$ based on the test functions $S_1(g,W)$ and $S_2(g,W)$, respectively, and the GELR test statistic $\hat{T}_{n,3}(\theta) := m_n^{-1/2}\hat{T}_n^{GELR}(\theta)$, we have*

(a) *Under Conditions 6, 7, 14, 8, and 9, for critical values $\hat{c}_F(\theta, 1-\alpha)$ obtained from plug-in asymptotics, the asymptotic size of the test based on $\hat{T}_{n,j}(\theta)$, $j = 1,2,3$ satisfies*

$$\liminf_n \inf_{(\theta,F)} P_F(T_n(\theta) \leq \hat{c}_F(\theta, 1-\alpha)) \geq 1-\alpha$$

(b) *Under assumptions 6-7, 8, 9, and 11, for critical values $\hat{c}_S(\theta.1-\alpha)$ obtained from subsampling, the asymptotic size of the test based on $\hat{T}_{n,j}(\theta)$, $j = 1,2,3$ satisfies*

$$\liminf_n \inf_{(\theta,F)} P_F(T_n(\theta) \leq \hat{c}_S(\theta, 1-\alpha)) \geq 1-\alpha$$

From Lemma 2 and Theorem 3, we can approximate the distribution of the QLR statistic arbitrarily well by a chi-bar square random variable, so that we can extend Proposition 3 to the case of non-Gaussian errors:

Corollary 3 *Suppose Condition 13 and the Assumptions of Theorem 3 hold, then the QLR statistic under the least favorable hypothesis, $\hat{T}_{nm_n}(\theta) = \sqrt{m_n}\min_{t\geq 0}(\zeta_n(\theta) - t)'\Omega_{nM}(\theta)^{-1}(\zeta_n(\theta) - t)$ converges in distribution to*

$$\frac{\hat{T}_{nm_n}(\theta) - \frac{m_n^{1/2}}{2}}{\sqrt{1+\bar{\sigma}_{m_n}(\theta)^2}} \xrightarrow{d} N(0,1)$$

Recall that Theorem 3 requires that $\frac{m_n^7}{n^2} \to 0$ which seems overly restrictive for deriving the distributions of quadratic forms of the type given by test functions S_1 and S_2.

We will now give an asymptotic normality result under the assumption that Ω is diagonal which sidesteps the argument from Theorem 3 and delivers a sharp restriction on the rate of m_n.

Proposition 4 *Suppose Condition 8 holds.*

(i) If $m_n \to \infty$ and $\frac{m_n}{n} \to 0$, and ζ_{mn} are strong α mixing with size $-\frac{r}{r-2}$, then for $S_1(g, \Omega^{-1})$ we have

$$\sqrt{m_n}\frac{2S_1(\varphi^{(1)}_{nm_n}(\hat{g}_n), \Omega_n^{-1}) - \mu_n}{2\sqrt{1+\bar{\sigma}_n^2}} \xrightarrow{d} N(0,1)$$

where the sequences μ_n and $\bar{\sigma}_n$ are defined in the appendix.

(ii) If $m_n \to \infty$ and $\frac{m_n}{n} \to 0$ and $\Omega_n(\theta)$ is diagonal, then the QLR statistic satisfies

$$\sqrt{m_n}\frac{2S_2(\varphi^{(1)}_{nm_n}(\hat{g}_n), \Omega_n^{-1}) - \mu_n}{\sqrt{1+\bar{\sigma}_n^2}} \xrightarrow{d} N(0,1)$$

where under the least favorable hypothesis $1 + \bar{\sigma}_n^2 = 5$.

where the rate condition $\frac{m_n}{n} \to 0$ is necessary for the conclusions.

This last result gives a *sharp* rate on the number of moment conditions for one relevant special case. Since for the subsample size, we have $\frac{b_n}{n} \to 0$, this rate result implies that for a range of rates m_n, subsampling fails whereas plug-in asymptotics remain valid. It is interesting to note that for this special case of the QLR statistic, the approximation error enters only through the mean of the censored censored moments, and, as one can see from the proof of Proposition 4, its magnitude depends mainly on the third cumulants of the marginal distributions of the components of $\xi_{in}(\theta)$.

If $\Omega_n(\theta)$ is not diagonal, the contributions of the individual components of the moment vector become interrelated through the projection implicit in the multivariate censoring problem, so the analogous argument would be more involved, and I leave this for future research. Also, this argument is specific to the quasi-likelihood ratio test, and does not extend to other convex test functions.

1.5 Discussion

In this paper I show how important insights from the literature on weak identification apply to set-identified problems. However, settings with moment inequalities differ from the standard GMM setup in that the shape of the identification region, which is the main object of interest, depends on which constraints are used for inference or estimation. In this sense, there are typically few or no "over-identifying" restrictions, and the sharp identified region can only be obtained if all available moment restrictions are used for estimation. Also, estimation and inference has to account for the presence of the slackness parameter which has the same dimension as the moment vector and can only be estimated conservatively as suggested by Chernozhukov, Hong, and Tamer (2007) and Andrews and Soares (2007).

My results on the rate of consistency also indicate that even though in many cases any finite number of constraints does not determine the sharp identification region, a set estimator using only a relatively small subset of moment inequalities may in fact be superior to a procedure based on a larger number of restrictions. In particular for the conditional moment inequality example, the approximation error discussed in Sections 2 and 3 decreases very fast even for small numbers of moment conditions, whereas the noise contribution is proportional to m_n. It would clearly be desirable to have a data-driven method to resolve this trade-off in practice, but this is beyond the scope of this paper.

The conditions needed for consistency of the set estimator may in practice be quite demanding, and we saw that unlike in some point-identified settings, inverse variance weighting does not lead to a weaker condition on the rates m_n and μ_n for the set estimator. In part, this is a result of the set estimator using a fixed critical value whereas the distribution of the criterion function is not asymptotically pivotal, but will typically vary across the parameter space. This suggests that for inference, parameter-dependent critical values should be used, especially in weakly identified settings, which has been the recommendation of the more recent literature.

The general inference result is also relatively demanding on the maximal number of moments compared to sample size. In particular, I show that we should expect approximations of the distribution using subsampling to be poor in particular if the distribution of the moment vector is asymmetric. If the number of moments is small relative to sample size, this leads only to a bias in the slackness parameters towards zero which makes inference conservative. However if m is large, subsampling also fails to approximate other features of the distribution, so that asymptotic size of the resulting confidence region may exceed the nominal level.

Chapter 2

Conditional Inference Procedures with Moment Inequalities

Inference on a finite-dimensional parameter in set-identified models is often subject to a number of moment inequalities that is significantly larger than the dimension of the parameter space. Whereas in a K-dimensional parameter space, at any given point on the boundary of the identified set, typically at most K population moment inequalities will be binding, an AR-type procedure will often test a much larger number of moments, even after applying a moment-selection procedure as in Andrews and Soares (2007). In the presence of a large number of moment restrictions, there are two factors which reduce the power of inference procedures: for one, the power of tests based on quadratic forms decreases in the number of degrees of freedom, and on the other hand, common tests for moment inequalities have to be conservative with respect to the slackness of non-binding constraints.

As for inference with moment equalities, the power of a chi-square type test decreases in the number of degrees of freedom for an Anderson-Rubin type statistics for a fixed noncentrality parameter (see e.g. Lehmann and Romano (2005), ch.14), and for the case of linear instrumental variables models with Gaussian errors, Andrews and Stock (2006) showed that AR type tests have only trivial limiting power under many weak moment asymptotics.

Also, since for partially identified problems, at any given point in the identification region, most moment inequalities are going to be slack. The relevant slackness parameter for the asymptotic distribution of the test statistic can only be estimated conservatively either by a moment selection procedure as suggested by Chernozhukov, Hong, and Tamer (2007), Bugni (2008) and Andrews and Soares (2007) for finitely many inequalities, or by preliminary estimation of a contact set if there is a continuum of conditions as in Chernozhukov, Lee, and Rosen (2008). In general, inference has to be conservative regardless of the procedure used to obtain critical values. For the AR-type statistic, the dimension of this nuisance parameter equals the number of moments used for inference. Therefore the problem of inference with moment inequalities exhibits an additional "curse of dimensionality" in that the critical values correspond to the least favorable value of a high-dimensional object.

In the weak instruments literature, there are two main approaches to eliminating variation in directions orthogonal to the parameter: Kleibergen (2002)'s LM test is based on the score of the concentrated objective function, whereas Moreira (2003)'s Conditional Likelihood Ratio (CLR) test conditions on a sufficent statistic for the nuisance parameter. The idea behind the LM test seems to adapt more readily to the GMM set-up (Kleibergen (2005)), but tests based on the LM statistic turn out to be dominated the conditional likelihood ratio test in power comparisons (Andrews, Moreira, and Stock (2006)). Also, the score may have multiple roots.

In this chapter, I propose an LM-type statistic that is based on lower-dimensional linear combinations of the original moment vector. This is an alternative to the Anderson-Rubin-type statistics several versions of which have been recommended for use in the recent literature (see e.g. Rosen (2008), Canay (2007), and Andrews and Jia (2008)). I will show that reducing the dimension of the moment vector can lead to a more powerful procedure.

Due to the geometry of the one-sided testing problem, the LM and CMD statistics proposed in this paper will in general not be asymptotically pivotal, so it will be necessary to obtain critical values from a simulation procedure.

A major concern in inference for over-identified models is that commonly used some testing procedures such as the AR test will yield empty confidence sets under mis-specification, or more generally that imposing moment constraints which do not hold at the population parameter leads to narrower confidence intervals, which may be mistaken for a greater precision of the inference procedure. A version of the proposed LM statistic will address this problem and guarantee non-empty confidence regions even if the moment inequality model is misspecified.

In the following section. I will define the general setup for inference considered in this paper, section three defines a Lagrange Multiplier (LM) statistic and a Conditional Minimum Distance (CMD) statistic for the moment inequalities, and section four discusses their theoretical asymptotic properties. Section five presents Monte Carlo simulation results comparing power of AR-, LM-, and CMD-tests. and the final section concludes.

2.1 Setup and Motivation

This paper considers inference for a K-dimensional parameter vector $\theta = (\theta_1, \dots, \theta_K)' \in \Theta$, a subset of $\mathbb{R}^K$. For a sample $Y_1, \dots, Y_n$ of i.i.d. observations, the population parameter θ_0 is assumed to satisfy M moment inequalities

$$\mathbb{E}[g(Y_i, \theta_0)] \geq 0 \tag{2.1}$$

where $g(y, \theta)$ is an M-dimensional function of y and θ with expectation $\bar{g}(\theta) := \mathbb{E}[g(Y_i, \theta)]$ and variance matrix $\Omega(\theta) := \mathbb{E}[(g(Y_i, \theta) - \bar{g}(\theta))(g(Y_i, \theta) - \bar{g}(\theta))']$. I will assume that the moment function $g(y, \theta)$ is bounded and twice continuously differentiable in θ for any value of y in the support of Y_i. For the purposes of my analysis, I also assume throughout that the number of moments M is greater than the dimension of the parameter vector, K.

Since the empirical restrictions imposed on the parameter are inequalities, in the present setting, the value of θ satisfying (2.1) will typically not be unique. The identification region

$$\Theta_I := \{\theta \in \Theta : \mathbb{E}[g(Y_i, \theta)] \geq 0\}$$

is defined as the set of parameter values for which all moment inequalities hold in the population, and any testing procedure based on these moment restrictions can (at best) only have nontrivial power against alternatives outside of the identification region.

Given the sample $Y_1, \dots, Y_n$, the continuously updated, inverse variance weighted GMM criterion is given by

$$\hat{Q}_n(\theta) := \min_{\nu \geq 0} (\hat{g}_n(\theta) - \nu)' \hat{\Omega}_n(\theta)(\hat{g}_n(\theta) - \nu)$$

where the M-dimensional vector $\hat{g}_n(\theta) := \frac{1}{n}\sum_{i=1}^n g(Y_i, \theta)$ is the sample moment, and $\hat{\Omega}_n(\theta)$ is a consistent estimator for the variance covariance matrix of $\hat{g}_n(\theta)$. Minimization of the quadratic form over non-negative values of $\nu \in \mathbb{R}_+^M$ should be understood as taking component-wise negative parts of $\hat{g}_n(\theta)$ in a way which takes into account the variance-covariance structure of the moment vector.

2.2 Test Statistics

The Anderson-Rubin (AR) type statistic for this estimation problem is given by the concentrated CUE objective function

$$AR_n(\theta) = n\hat{Q}_n(\theta) := \min_{\nu \geq 0} (\hat{g}_n(\theta) - \nu)' \hat{\Omega}(\theta)^{-1} (\bar{g}_n(\theta) - \nu)$$

In the case of moment equalities, inference based on the AR statistic may yield empty confidence intervals with nonzero probability (see Kleibergen (2002) and Kleibergen (2005)), and moreover for a large number of degrees of freedom, the test may have low power. In fact, for the point-identified linear IV model, which is a special case of our model up to the restrictions on the growth rate of the number of moments, Andrews and Stock (2006) showed that under many weak moment asymptotics, the AR test has asymptotic power equal to size.

2.2.1 Score (LM) Statistic

Concentrating out the slackness parameters ν, the continuously updated (CU) inverse variance weighted criterion function for this problem is given by

$$Q_n(\theta) = n \min_{\nu \geq 0} (\hat{g}_n(\theta) - \nu)' \hat{\Omega}_n(\theta)^{-1} (\hat{g}_n(\theta) - \nu) = n(\hat{g}_n(\theta) - \hat{\nu}^*(\theta))' \hat{\Omega}(\theta)^{-1} (\hat{g}_n(\theta) - \hat{\nu}^*(\theta))$$

where for positive definite $\hat{\Omega}_n(\theta)$ $\hat{\nu}^*(\theta) := \arg\min_{\nu \geq 0} (\hat{g}_n(\theta) - \nu)' \hat{\Omega}_n(\theta)^{-1} (\hat{g}_n(\theta) - \nu)$ exists and is unique, see e.g. Luenberger (1969).

Note that even under the null hypothesis and at a parameter value θ in the identification region, the expectation of the moment vector will be non-negative expectation, but not necessarily equal to zero, so that it is important to re-center the moment vector using an estimator for the population mean $\bar{g}(\theta) := \mathbb{E}[g(Y_i, \theta)]$. In order to keep the problem numerically tractable as well as make sure that the limiting distribution of the relevant functions of the estimated variance matrix is continuous, we will impose non-negativity of the moment functions by using a suitable component-wise transformation of the sample mean depending on sample size, defined as

$$\bar{\psi}_{1nm}(\theta) := \psi_{1n}(\hat{g}_n(\theta)) = \begin{cases} \hat{g}_{mn}(\theta) & \text{if } \eta_n \frac{\hat{g}_{mn}(\theta)}{\hat{\omega}^*_{mm}(\theta)} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

for the mth component of $\hat{g}_n(\theta)$, where $\hat{\omega}^*_{mm}(\theta) := \sqrt{\widehat{\mathrm{Var}}(\hat{g}_{mn}(\theta))}$, and η_n is a sequence of nonnegative numbers such that

$$P\left(\eta_n \to 0 \text{ and } \liminf_n \sqrt{\frac{n}{2 \log \log n}} \eta_n > 1\right) = 1$$

For i.i.d. data, we will estimate the variance-covariance matrix of the moment vector with

$$\hat{\Omega}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} (g(Y_i, \theta) - \bar{\psi}_{1n}(\theta))(g(Y_i, \theta) - \bar{\psi}_{1n}(\theta))'$$

Denoting

$$\hat{G}_{nk}(\theta) := \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_k}g(Y_i,\theta), \qquad \hat{G}_n(\theta) := [\hat{G}_{n1}(\theta)',\dots,\hat{G}_{nK}(\theta)']'$$

$$\hat{C}_{nk}(\theta) := \frac{1}{n}\sum_{i=1}^{n}(g(Y_i,\theta) - \bar{\psi}_{1n}(\theta))'\frac{\partial}{\partial\theta_k}g(Y_i(\theta)$$

and the stacked matrix $\hat{C}_n(\theta) = [\hat{C}_{n1}(\theta)',\dots,\hat{C}_{nK}(\theta)']'$, we can define

$$\tilde{G}_n(\theta) := \hat{G}_n(\theta) - (\iota_K' \otimes ((\hat{g}_n(\theta) - \bar{\psi}_{1n}(\theta))'\hat{\Omega}_n(\theta)^{-1}))\hat{C}_n(\theta)$$

At any minimizer θ of the concentrated criterion, we then have the following first-order condition on the subgradient of $Q_n^*(\theta)$

$$\hat{D}_n^*(\theta) := \nabla_\theta Q_n^*(\theta) = \tilde{G}_n^*(\theta)'\hat{\Omega}_n(\theta)^{-1}(\hat{g}_n(\theta) - \nu^*(\theta)) = 0 \tag{2.2}$$

where $\tilde{G}_n^*(\theta) = \tilde{G}_n(\theta)\hat{\Psi}_n(\theta)$ and the selection matrix $\hat{\Psi}_n(\theta)$ is a diagonal matrix for which the mth diagonal element is an indicator $\hat{\Psi}_{nmm}(\theta) = \mathbb{1}\{\hat{\nu}_{nm}^*(\theta) > 0\}$.

Even though $\nu^*(\theta) \geq 0$, the coefficients on the linear combinations are not guaranteed to have non-negative coefficients, so that 2.2 will in general not give informative inequality restrictions. Instead we are going to define

$$\hat{D}_{n+}(\theta) := \tilde{G}_{n,|\cdot|}(\theta)'\hat{\Omega}_n(\theta)^{-1}(\hat{g}_n(\theta) - \bar{\psi}_{2n}(\theta)) \tag{2.3}$$

where the columns of $\tilde{G}_{n,|\cdot|}$ are defined by

$$\tilde{G}_{mn,|\cdot|} := \tilde{G}_{mn,+} - \tilde{G}_{mn,-} := \text{Proj}(\tilde{G}_{mn}^*|\mathbb{R}_+^m, \hat{\Omega}_n^{-1}) - \text{Proj}(\tilde{G}_{mn}^*|\mathcal{C}_{\hat{\Omega}^{-1}}^0, \hat{\Omega}_n^{-1})$$

where $\text{Proj}(x|\mathcal{C},W) := \arg\min_{t\in\mathcal{C}}(x-t)'W(x-t)$ denotes the orthogonal projection of the vector x onto the cone $\mathcal{C}$ with respect to the inner product defined by the weighting matrix W, and $\mathcal{C}_W^0 := \{y \in \mathbb{R}^m : y'Wx \leq 0 \text{ for all } x \in \mathbb{R}_+^m\}$ is the polar cone to the positive orthant of $\mathbb{R}^m$ with respect to the inner product defined by the weighting matrix W. The function $\bar{\psi}_{mn}(\theta) := \psi_2(\hat{g}_n(\theta))$ is a component-wise transformation of the sample moment vector and does not necessarily have to be the same as the moment selection function $\psi_{1n}(\cdot)$. This allows in principle to extend this analysis to a refined moment selection procedure as in Andrews and Jia (2008).

We are now going to define a statistic based on the modified score from equation 2.3. The modified LM statistic for the inequality testing problem is given by

$$LM(\theta) := \min_{t\in\mathbb{R}^k_+}(\hat{D}_{n+}(\theta) - t)' \left(\tilde{G}_{n,|\cdot|}(\theta)'\hat{\Omega}_n(\theta)^{-1}\tilde{G}_{n,|\cdot|}(\theta)\right)^{-1} (\hat{D}_{n+}(\theta) - t) \tag{2.4}$$

Note that whenever $\bar{g}(\theta) \geq 0$, by construction $\tilde{G}_{n,|\cdot|}(\theta)'\hat{\Omega}_n(\theta)'\bar{g}(\theta) \geq 0$. We will show below that under regularity conditions $\hat{g}_n(\theta) - \bar{g}(\theta)$ and $\tilde{G}_{n,|\cdot|}(\theta)$ are asymptotically independent at any value of for which $\bar{g}(\theta_0) \geq 0$, so that for any $\theta_0 \in \Theta_I$ in the identification region, $\mathbb{E}[\hat{D}_{n+}(\theta_0)|\tilde{G}_{n,|\cdot|}(\theta_0)] \geq 0$. Therefore, the non-negativity restrictions on the slackness parameters for the original moment inequalities imply non-trivial restrictions on the linear combinations of slackness parameters in the modified score.

The LM statistic as defined in 2.4 only makes use of a number of linear combinations of the sample moments equal to the dimension of the parameter vector. Hence, in the case of linear moment restrictions there exists at least one parameter value θ^* at which all linear combinations of moment inequalities hold as equalities, so that at this value, $LM(\theta^*) = 0$. This ensures that in the linear case, a confidence interval based on the LM statistic is non-empty with probability one regardless of whether the model is correctly specified.

Note also that using the modified score of the CU objective function for inference reduces the computational cost of simulating critical values for the statistic since the quadratic optimization problem we have to solve in order to obtain (2.4) is over a vector of the same dimension as the moment vector used for inference.

Aggregating moment inequalities will likely result in loss of power against certain alternatives because binding moments at parameters outside the identification region may cancel against restrictions that are slack at that parameter value. That means that even though under a given alternative $\theta_A \notin \Theta_0$ some moments will not be satisfied, i.e. $\bar{g}(\theta_A) \ngeq 0$, the linear combination $\hat{D}(\theta_A)'\hat{\Omega}(\theta_A)^{-1}\bar{g}(\theta_A)$ may still be non-negative. Even when inference can only be conservative, an optimal procedure should therefore let the dimension of the linear combinations to be larger than k and increase with sample size.

This stands in contrast to point-identified GMM problems for which typically all locally identifying information contained in the moment restrictions can be aggregated into appropriately chosen linear combinations of dimension equal to the number of parameters. The problem with the aggregation for the moment inequalities can be interpreted as an instance of the lack of "global" power of the Kleibergen test against "irrelevant alternatives", i.e. roots of the score which do not correspond to a minimum of the CUE objective function.

2.2.2 Conditional Minimum Distance (CMD) Statistic

In order to address the potential lack of power of score type tests against points corresponding to other extrema of $\hat{Q}_n(\theta)$, I will now propose an alternative statistic which retains some of the potential advantages of the (conditional) LM statistic while improving its power against extraneous roots of the score.

In GMM inference with moment equalities, a generalization of Moreira (2003)s Conditional Likelihood Ratio (CLR) test can be represented as a weighted combination of the AR and the LM statistic, using a test statistic for the rank of the average Jacobian of the moment functions as weights. Analogous to Kleibergen (2005), we can define the following conditional minimum distance (CMD) statistic

$$CMD_n(\theta) = \frac{1}{2}\left\{AR_n(\theta) - RK_n(\theta) + \sqrt{(AR_n(\theta) - RK_n(\theta))^2 + 4RK_n(\theta)LM_n(\theta)}\right\} \tag{2.5}$$

where $RK_n(\theta) := \tilde{G}_n(\theta)'\hat{\Omega}(\theta)^{-1}\tilde{G}_n(\theta)$ is a statistic for a test of the rank condition. For the case of moment equalities, Kleibergen (2005) showed that this modification removes the extraneous roots of the score equation for the CUE estimator.

For moment equalities, the conditional minimum distance statistic should be expected to improve power over an Anderson-Rubin type procedure because subtracting off the minimized value of the statistic over the parameter space reduces the degrees of freedom by the number of over-identifying restrictions, and conditioning on $RK_n(\theta)$ eliminates the nuisance parameter corresponding to the Jacobians from the asymptotic distribution of the statistic. In the next section, $RK_n(\theta)$ will be shown to be asymptotically independent of $AR_n(\theta)$ and $LM_n(\theta)$. However, since we have to impose non-negativity on the coefficients for linear combinations of moments, it will in general not be possible to decompose $AR_n(\theta)$ into a function of the LM statistic and a component independent of $LM_n(\theta)$, so that in order to obtain asymptotically valid critical values for a CMD test, $AR_n(\theta)$ and $LM_n(\theta)$ have to be simulated jointly.

2.3 Large Sample Theory

I will now give results on the asymptotic distribution of the statistics $LM_n(\theta)$ and $CMD_n(\theta)$ which justify the simulation procedures to obtain critical values proposed in the previous section. In the following, I will denote the population expectations of the moment vector and its Jacobian by

$$\bar{g}(\theta) := \mathbb{E}[g(Y_i, \theta)], \quad \bar{G}_{nk}(\theta) := \mathbb{E}\left[\frac{\partial}{\partial\theta_k} g(Y_i, \theta)\right], \quad \text{and } \bar{G}_n(\theta) := [\bar{G}_{n1}(\theta)', \dots, \bar{G}_{nK}(\theta)']'$$

respectively.

Assumption 7 *(i) The sample moment $\hat{g}_n(\theta)$ and its first derivative $\hat{G}_n(\theta)$ satisfy a central limit theorem,*

$$\sqrt{n}\begin{pmatrix} \hat{g}_n(\theta) - \bar{g}(\theta) \\ \mathrm{vec}(\hat{G}_n(\theta) - \bar{G}(\theta)) \end{pmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Omega(\theta) & C(\theta) \\ C(\theta) & V(\theta) \end{bmatrix}\right)$$

uniformly in θ, where (ii) $\Omega(\theta)$ is positive definite, and $\begin{bmatrix} \Omega(\theta) & C(\theta) \\ C(\theta) & V(\theta) \end{bmatrix}$ *is positive semi-definite for all θ.*

Assumption 7 (i) can be replaced by relatively standard lower-level assumptions - e.g. the moment functions and their Jacobian being bounded Lipschitz - for a summary see e.g. van der Vaart (1998) or van der Vaart and Wellner (1996). As in Kleibergen (2005), the requirement that the joint variance matrix of the moments has to be only positive semi-definite also accommodates the practically relevant case in which elements of the Jacobian are non-random.

Assumption 8 *(i) For the estimator of Ω we have that uniformly in θ,*

$$\hat{\Omega}_n^*(\theta) := \frac{1}{n}\sum_{i=1}^{n}(g(Y_i,\theta) - \hat{g}_n(\theta))(g(Y_i,\theta) - \hat{g}_n(\theta))' \xrightarrow{p} \Omega(\theta)$$

(ii) Uniformly in θ,

$$\hat{C}_{nk}^*(\theta) := \frac{1}{n}\sum_{i=1}^{n}(g(Y_i,\theta) - \hat{g}_n(\theta))(G(Y_i,\theta) - \hat{G}_n(\theta))' \xrightarrow{p} C_k(\theta)$$

The first two parts of Assumption 8 require that the covariance matrix of the moment functions and Jacobians can be estimated consistently for i.i.d. data, which is true under commonly imposed regularity conditions (e.g. existence of fourth moments).

Assumption 9 *At $\theta \in \Theta_I$ for each moment m, we have either (i) $\bar{g}_{mn}(\theta) \to g > 0$, or (ii) $(n \log\log n)^{1/2}\bar{g}_{mn}(\theta) \to 0$.*

Assumption 9 restricts the behavior of the slackness of population moments in the population in the identification region. Case (i) corresponds to conventional strong moments with fixed parameters under the null hypothesis, and (ii) covers cases of weak identification and near-binding moments.

As shown by Rosen (2008), under Assumptions 7 and 8, $AR_n(\theta)$ converges in distribution to a chi-bar squared distribution with M degrees of freedom,

$$AR_n(\theta) \xrightarrow{d} \min_{t\in\mathbb{R}_+^M}(\bar{g}(\theta) + Z - t)'\Omega(\theta)^{-1}(\bar{g}(\theta) + Z - t) =: \bar{\chi}^2(\bar{g}(\theta), \Omega(\theta), \mathcal{C}_{\Omega(\theta)^{-1}})$$

where $Z \sim N(0, I_M)$.

In order to analyze the asymptotic properties of the score test statistic defined in (2.4), let us first consider the joint distribution of the sample moment and the estimated Jacobian, $\tilde{G}_n(\theta)$:

Proposition 5 *Under Assumptions 7 and 8, as $n \to \infty$,*

$$\sqrt{n}\begin{pmatrix} \hat{g}_n(\theta) - \bar{g}_n(\theta) \\ \mathrm{vec}(\tilde{G}_n(\theta) - \bar{G}_n(\theta)) \end{pmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Omega(\theta) & 0 \\ 0 & V(\theta) - C(\theta)'\Omega(\theta)^{-1}C(\theta) \end{bmatrix}\right)$$

In particular, $\hat{g}_n(\theta)$ and $\tilde{G}_n(\theta)$ are asymptotically independent which implies that $\tilde{G}_{n,|\cdot|}(\theta)$ is also asymptotically independent of $\hat{g}_n(\theta)$, so that from Slutsky's theorem and a central limit theorem for $\hat{g}_n(\theta)$ we can derive the asymptotic distribution of the modified score, $\hat{D}_{n+}(\theta)$ conditional on $\tilde{G}_n(\theta)$:

Corollary 4 *Under the Assumptions of Proposition 5,*

$$S_n(\theta) := \sqrt{n}\left(\tilde{G}_{n,|\cdot|}(\theta)'\hat{\Omega}_n(\theta)^{-1}\hat{G}_{n,|\cdot|}(\theta)\right)^{-1}\tilde{G}_{n,|\cdot|}(\theta)'\hat{\Omega}_n(\theta)^{-1}(\hat{g}_n(\theta) - \bar{g}(\theta)) \rightsquigarrow \psi(\theta)$$

where conditional on $\tilde{G}_{n,|\cdot|}(\theta)$, $\psi(\theta) \sim N(0, I)$.

It follows that the asymptotic distribution for the LM statistic defined in 2.4 is given by

Corollary 5 *Under the Assumptions of Proposition 5, the asymptotic distribution for the modified LM statistic is given by*

$$\begin{aligned} LM(\theta) &= \min_{t\in\mathbb{R}^k_+}(\hat{D}_{n+}(\theta) - t)'\left(\tilde{G}_{n,|\cdot|}(\theta)'\hat{\Omega}_n(\theta)^{-1}\hat{G}_{n,|\cdot|}(\theta)\right)^{-1}(\hat{D}_{n+}(\theta) - t) \\ &\xrightarrow{d} \bar{\chi}^2(\bar{g}, \mathcal{C}_{\hat{W}(\theta)^{-1}}, \hat{W}(\theta)^{-1}) := \min_{\nu\geq 0}(Z - \nu)'\hat{W}^{-1}(Z - \nu) \text{ for } Z \sim N(\mu, \Omega) \end{aligned}$$

conditional on $\hat{D}(\theta)$, where $\hat{W}(\theta) = \tilde{G}_{n,|\cdot|}(\theta)'\hat{\Omega}_n(\theta)^{-1}\hat{G}_{n,|\cdot|}(\theta)$, and $\bar{\chi}^2(\mu, \Omega, \mathcal{C}_W)$.

Hence, inference based on the pseudo-LM statistic again reduces to a chi-bar-square testing problem where critical values can be obtained by simulation given the estimates $\tilde{G}_{n,|\cdot|}(\theta)$ and $\hat{\Omega}_n(\theta)$. By the conditioning argument we can replace the population expectation of the projected Jacobians with their sample analogs without having to adjust the asymptotic distribution for the fact that they are estimated from the same data as the sample moments. It should be pointed out that this argument is only valid for the case of a finite-dimensional moment vector.

2.4 Simulations

In this subsection, we compare the power functions of tests based on the AR-type and LM-type statistic for the linear model. More specifically, we generate data from

$$y_i^* = x_i\beta + \varepsilon$$
$$x_i = z_i\pi + \nu$$

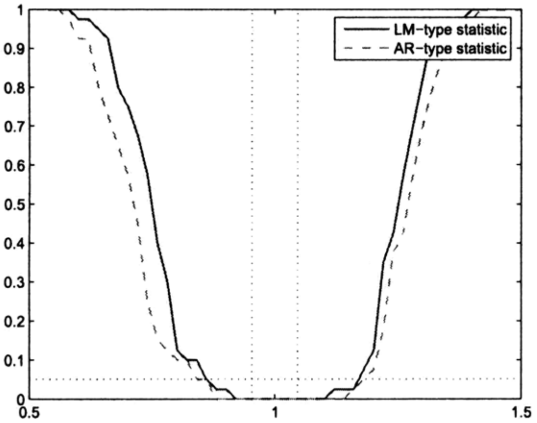
where $(\varepsilon, \nu) \sim N\left(0, \sigma^2 \begin{bmatrix} 1 & \varrho \\ \varrho & 1 \end{bmatrix}\right)$ and z_i is an M-dimensional nonnegative random vector with unit variance which is independent of (ε, ν). For estimation, we assume that we do not observe y_i^*, but bounds such that $y_{il} \leq y_i^* \leq y_{iu}$ with probability one, and $\mathbb{E}[y_u - y_l|z] = h$ for some positive constant h. We then form moment functions

$$g_{1i}(\beta) = z_i(y_{iu} - x_i\beta), \qquad g_2(\beta) = -z_i(y_{il} - x_i\beta)$$

The graphs show the simulated rejection probabilities of the AR and LM type tests at a nominal 5% significance level for different values of β using critical values obtained by simulation from a Gaussian distribution under the least favorable hypothesis. The data was generated under $\beta_0 = 1$, and the boundaries of the population identification region for a particular choice of parameters are plotted as vertical dotted lines.

The "first stage" parameter π was chosen to be small in all scenarios so that generalized moment selection would not have been likely to detect any slack moments for the range of hypotheses on β considered in the simulation study. From the power functions, we can see that the rejection probabilities are less than 5% for parameter values in the identification region (marked by the vertical dotted lines in the graphs), indicating that both testing procedures are conservative, and have confidence size

Figure 2-1: Power Comparison between AR and LM Type Test

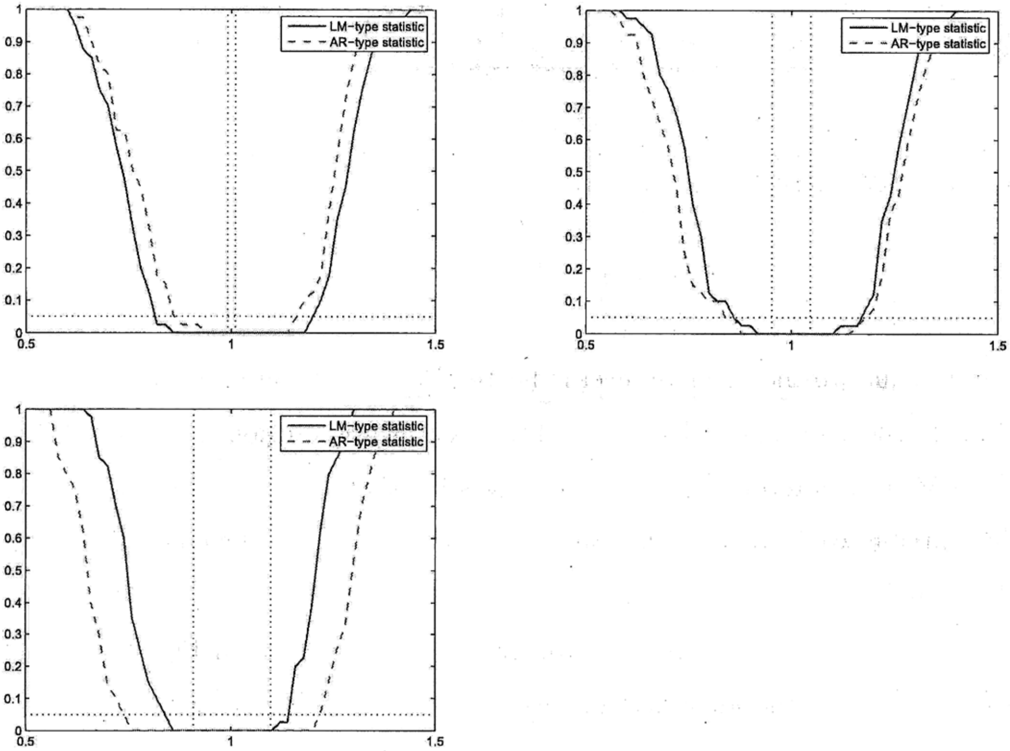


Simulations with $N = 1000, M = 20, \|\pi\| = \sqrt{\frac{60}{1000}}, \varrho = 0.3, h = 0.05, \sigma = 2$.

less than or equal to the nominal level indicated by the horizontal dotted line. Most notably, at least in a neighborhood around the identification region, the LM test dominates the AR test in terms of power, also suggesting that confidence regions based on the LM statistic would be considerably smaller than those constructed by inverting the AR statistic.

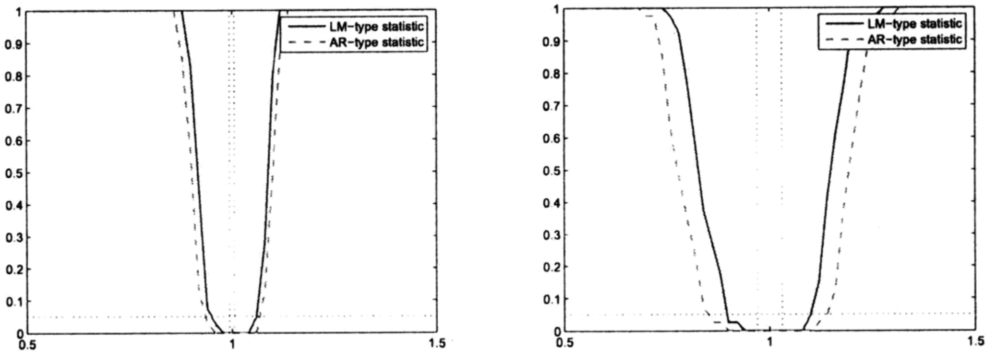
The simulations in Figure 2 are based on the same scenario as in Figure 1, except that the diameter of the identification region varies from very short ($h = 0.02$) to relatively wide ($h = 0.2$). The simulations show that the LM test dominates the AR test except for a very narrow identification region, in which case at each boundary of Θ_I, the moment selection procedure fails to detect some of the slack inequalities which correspond to the other boundary point of the set. As Figure 3 shows, this problem becomes less relevant if the number of moment conditions increases, because in this case the benefits from reducing the number of degrees of freedom of the procedure seems to outweigh the potential power loss from aggregating the moment conditions. In particular, the relative performance of the score-type test statistic seems to improve as we consider testing problems with a larger number of moment conditions.

Figure 2-2: Power of AR and LM Test for $M = 20$ and Different Lengths of the Identification Region



Simulations with $N = 1000, M = 20, \varrho = 0.3, \|\pi\| = \sqrt{\frac{60}{1000}}, \sigma = 2$, and h varying from $h = 0.02$(top left) to $h = 0.1$ (top right), and $h = 0.2$(bottom).

Figure 2-3: Power of AR and LM Test for $M = 50$ and Different Lengths of the Identification Region



Simulations with $N = 1000, M = 50, \varrho = 0.3, \|\pi\| = \sqrt{\frac{60}{1000}}, \sigma = 2$, and h taking values $h = 0.02$(left) and $h = 0.1$ (center).

2.5 Discussion

The power comparisons between AR-type and the pseudo-LM statistic in section 5 indicate that especially under weak identification, aiming directly at the sharp identification region need not necessarily give the smallest confidence sets, but taking suitable linear combinations of moments together with a conditioning argument can enhance the power of inference procedures under reasonable assumptions. The combination of a conditioning argument with moment selection as in the definition of the CMD statistic looks very promising, but finding the optimal combination of these two aspects and systematic power comparisons with alternative procedures are beyond the scope of this paper and will be left for future research. However not all recommendations are as clear-cut due to the inherent "second-best" nature of one-sided testing problems that has long been known in the literature.

The simulation results suggest that for a range of practically relevant settings, the proposed new statistics dominate the AR type procedures which are recommended and used widely in the literature on moment inequalities. Since under regularity conditions, the GELR class of test statistics is asymptotically equivalent to the QLR/AR test statistic, any test from that class should also be expected to inherit the same drawbacks. It should be pointed out that this can be reconciled with Canay (2007)'s large deviations optimality result for the ELR statistic as follows: large-deviations optimality only means that for any choice of a critical value, an ELR hypothesis test solves the trade-off between type-I and type-II error optimally in the limit. Since the asymptotic distribution of the ELR test for moment inequalities depends on a nuisance parameter which can't be estimated consistently, estimated critical values are conservative - i.e. it is in general not possible to control size precisely. Therefore, large-deviations optimality does not imply that a feasible size α test based on the ELR statistic is more powerful than alternative procedures. Furthermore, the arguments behind Kleibergen (2005)'s LM statistic and Moreira (2003)'s CLR test involve a conditioning argument whereas the large-deviations optimality result is on unconditional inference.

In classical GMM problems, the CMD test outperforms the AR test in part because it conditions on a sufficient statistic for the Jacobians of the moment functions, which are a potentially high-dimensional nuisance parameter for inference on the parameter of interest. However, for one-sided GMM-type testing problems, the slackness parameters of the moment inequalities introduce an additional nuisance parameter which can't be eliminated in a similar fashion. Therefore, it looks promising to combine the use of a CMD or LM-type statistic with a refined moment selection procedure as in Andrews and Jia (2008) using a tuning parameter which remains finite. Working out this connection properly is beyond the scope of this paper and will be left for future research.

Chapter 3

Inference on Sets in Finance

Joint with Victor Chernozhukov and Emre Kocatulum

3.1 Introduction

In this paper we introduce various set inference problems as they appear in finance and propose practical and powerful inferential tools. Our tools will be applicable to any problem where the set of interest solves a system of estimable inequalities, though we will particularly focus on the following two problems: The first problem will deal with mean-variance sets of stochastic discount factors and the second with mean-variance sets of admissible portfolios.

Let us now introduce the problem. We begin by recalling two equations used by Cochrane (2005) to effectively summarize the theory of asset pricing:

$$\begin{aligned} P_t &= E_t[M_{t+1}X_{t+1}] \\ M_{t+1} &= f(Z_{t+1}, parameters), \end{aligned}$$

where P_t is an asset price, X_{t+1} is the asset payoff, M_{t+1} is the stochastic discount factor (SDF) or pricing kernel (PK), which is a function f of some data Z_{t+1} and parameters, and E_t is the conditional expectation given information at time t. The set of SDFs M_t that can price existing assets generally form a proper set, that is, a set that is not a singleton. SDFs are not unique, because the existing payoffs to assets do not span the entire universe of possible random payoffs. Dynamic asset pricing models provide families of potential SDFs, for example, the standard consumption model predicts that an appropriate SDF can be stated in terms of intertemporal marginal rate of substitution:

$$M_t = \beta \frac{u'(C_{t+1})}{u'(C_t)},$$

where u denotes a utility function parameterized by some parameters, C_t denotes consumption at time t, and β denotes the subjective discount factor.

The basic econometric problem is to check which families of SDFs price the assets correctly and which do not. In other words, we want to check whether given families or subfamilies of SDFs are valid or not. One leading approach for performing the check is to see whether mean and standard deviation of SDFs

$$\{\mu_M, \sigma_M\}$$

are admissible. The set of admissible means and standard deviations

$$\Theta_0 := \{ \text{ admissible pairs } (\mu, \sigma^2) \in R^2 \cap K\},$$

which is introduced by Hansen and Jagannathan (1991) is known as the Hansen-Jagannathan set and the boundary of the set Θ_0 is known as the Hansen-Jagannathan bound. In order to give a very specific, canonical example, let v and Σ denote the vector of mean returns and covariance matrix to assets $1, ..., N$ which are assumed not to vary with information sets at each period t. Let us denote

$$A = v'\Sigma^{-1}v, B = v'\Sigma^{-1}1_N, C = 1_N'\Sigma^{-1}1_N \tag{3.1}$$

where 1_N is a column vector of ones. Then the minimum variance $\sigma^2(\mu)$ achievable by a SDF given mean μ of the SDF is equal to

$$\sigma^2(\mu) = (1-\mu v)'\Sigma^{-1}(1-\mu v) = A\mu^2 - 2B\mu + C$$

Therefore, the HJ set is equal to

$$\Theta_0 = \{\underbrace{(\mu,\sigma)}_{\theta} \in \underbrace{\mathbb{R}^2 \cap K}_{\Theta} : \underbrace{\sigma(\mu) - \sigma}_{m(\theta)} \leq 0\},$$

where K is any compact set. That is,

$$\Theta_0 = \{\theta \in \Theta : m(\theta) \leq 0\}.$$

Note that the inequality-generating function $m(\theta)$ depends on the unknown parameters, the means and covariance of returns, $m(\theta) = m(\theta, \gamma)$ and $\gamma =$ vec (v, Σ).

Let us now describe the second problem. The classical Markowitz (1952) problem is to minimize the risk of a portfolio given some attainable level of return:

$$\min_w E_t[r_{p,t+1} - E_t[r_{p,t+1}]]^2 \text{ such that } E_t[r_{p,t+1}] = \mu,$$

where $r_{p,t+1}$ is portfolios return, determined as $r_{p,t+1} = wr_{t+1}$, where w is a vector of portfolio "weights" and r_{t+1} is a vector of returns on available assets. In a canonical version of the problem, we have that the vector of mean returns v and covariance of returns Σ do not vary with time period t, so that the problem becomes:

$$\sigma(\mu) = \min_w w'\Sigma w \text{ such that } w'v = \mu.$$

An explicit solution for $\sigma(\mu)$ takes the form,

$$\sigma^2(\mu) = \frac{C\mu^2 - 2B\mu + A}{AC - B^2}$$

where A,B and C are as in equation 3.1.

Therefore, the Markowitz (M) set of admissible standard deviations and means is given by

$$\Theta_0 = \{\underbrace{(\mu, \sigma)}_{\theta} \in \underbrace{\mathbb{R}^2 \cap K}_{\Theta} : \underbrace{\sigma(\mu) - \sigma}_{m(\theta)} \leq 0\},$$

that is,

$$\Theta_0 = \{\theta \in \Theta : m(\theta) \leq 0\}.$$

The boundary of the set Θ_0 is known as the efficient frontier. Note that as in HJ example, the inequality-generating function $m(\theta)$ depends on the unknown parameters, the means and covariance of returns, $m(\theta) = m(\theta, \gamma)$, where $\gamma = \text{ vec } (v, \Sigma)$.

The basic problem of this paper is to develop inference methods on HJ and M sets, accounting for uncertainty in the estimation of parameters of the inequality-generating functions. The problem is to construct a confidence region R such that

$$\lim_{n \to \infty} P\{\Theta_0 \subseteq R\} = 1 - \alpha.$$

We will construct confidence regions for HJ sets using LR and Wald-type Statistics, building on and simultaneously enriching the approaches suggested in Chernozhukov, Hong, and Tamer (2007), Beresteanu and Molinari (2008), and Molchanov (1998). We also would like to ensure that confidence regions R are as small as possible and converge to Θ_0 at the most rapid attainable speed. We need the confidence region R for entire set Θ_0 in order to test validity of sets of SDFs. Once R is constructed, we can test infinite number of composite hypotheses, current and future, without compromising the significance level. Indeed, a typical application of HJ sets determines which sets of (μ, σ)'s within a given family fall in the HJ set and which do not. Similar comments about applicability of our approach go through for the M sets as well.

Our approach to inference using weighted Wald-type statistics complements and enriches the approach based on the directed Hausdorff distance suggested in Beresteanu and Molinari (2008) and Molchanov (1998). By using weighting in the construction of the Wald-type statistics, we endow this approach with better invariance properties to parameter transformations, which results in noticeably sharper confidence sets, at least in the canonical emprical example that we will show. Thus, our construction is of independent interest for this type of inference, and is a useful complement to the work of Beresteanu and Molinari (2008) and Molchanov (1998). Furthermore, our results on formal validity of the bootstrap for LR-type and W-type statistics are also of independent interest.

The rest of the paper is organized as follows. In Section 2 we present our estimation and inference results. In Section 3 we present an empirical example, illustrating the constructions of confidence sets for HJ sets. In Section 4 we draw conclusions and provide direction for further research. In the Appendix, we collect the proofs of the main results.

3.2 Estimation and Inference Results

3.2.1 Basic Constructions

We first introduce our basic framework. We have an inequality-generating function:

$$m : \Theta \mapsto \mathbb{R}.$$

The set of interest is the solution of the inequalities generated by the function $m(\theta)$ over a compact parameter space Θ:

$$\Theta_0 = \{\theta \in \Theta : m(\theta) \leq 0\}.$$

A natural estimator of Θ_0 is its empirical analog

$$\widehat{\Theta}_0 = \{\theta \in \Theta : \widehat{m}(\theta) \leq 0\}.$$

where $\widehat{m}(\theta)$ is the estimate of the inequality-generating function. For example, in HJ and M examples, the estimate takes the form

$$\hat{m}(\theta) = m(\theta, \hat{\gamma}), \quad \hat{\gamma} = \text{ vec } (\widehat{v}, \widehat{\Sigma}).$$

Our proposals for confidence regions are based on (1) LR-type statistic and (2) Wald-type statistic. The LR-based confidence region is

$$R_{LR} = \left\{\theta \in \Theta : \left[\sqrt{n}\hat{m}(\theta)/s(\theta)\right]_+^2 \leq \hat{k}(1-\alpha)\right\}, \tag{3.2}$$

where $s(\theta)$ is the weighting function; ideally, the standard error of $\hat{m}(\theta)$; and $\hat{k}(1-\alpha)$ is a suitable estimate of

$$k(1-\alpha) = (1-\alpha) - \text{ quantile of } \mathcal{L}_n,$$

where

$$\mathcal{L}_n = \sup_{\theta \in \Theta_0} \left[\sqrt{n}\hat{m}(\theta)/s(\theta)\right]_+^2 \tag{3.3}$$

is the LR-type statistic, as in Chernozhukov, Hong, and Tamer (2007).

Our Wald-based confidence region is

$$R_W = \{\theta \in \Theta : [\sqrt{n}d(\theta, \widehat{\Theta}_0)/w(\theta)]^2 \leq \hat{k}(1-\alpha)\}, \tag{3.4}$$

where $w(\theta)$ is the weighting function, particular forms of which we will suggest later; and $\hat{k}$ is a suitable estimate of

$$k(1-\alpha) = (1-\alpha) - \text{ quantile of } \mathcal{W}_n,$$

where $\mathcal{W}_n$ is the weighted W-statistic

$$\mathcal{W}_n = \sup_{\theta \in \Theta_0} [\sqrt{n}d(\theta, \widehat{\Theta}_0)/w(\theta)]^2. \tag{3.5}$$

Recall that quantity $d(\theta, \widehat{\Theta}_0)$ is the distance of a point θ to a set $\widehat{\Theta}_0$, that is,

$$d(\theta, \widehat{\Theta}_0) := \inf_{\theta' \in \widehat{\Theta}_0} \|\theta - \theta'\|.$$

In the special case, where the weight function is flat, namely $w(\theta) = w$ for all θ, the W-statistic $\mathcal{W}_n$ becomes the canonical directed Hausdorff distance (Molchanov (1998), Beresteanu and Molinari (2008)):

$$\sqrt{\mathcal{W}_n} \propto d(\Theta_0, \widehat{\Theta}_0) = \sup_{\theta \in \Theta_0} \inf_{\theta' \in \widehat{\Theta}_0} \|\theta - \theta'\|.$$

The weighted statistic (3.5) is generally *not* a distance, but we argue that it provides a very useful extension of the canonical directed Hausdorff distance. In fact, in our empirical example precision weighting dramatically improves the confidence regions.

3.2.2 A Basic Limit Theorem for LR and W statistics

In this subsection, we develop a basic result on the limit laws of the LR and W statistics. We will develop this result under the following general regularity conditions:

R.1 The estimates $\theta \mapsto \widehat{m}(\theta)$ of the inequality-generating function $\theta \mapsto m(\theta)$ are asymptotically Gaussian, namely, we have that in the metric space of bounded functions $\ell^\infty(\Theta)$

$$\sqrt{n}(\widehat{m}(\theta) - m(\theta)) =_d G(\theta) + o_P(1),$$

where $G(\theta)$ is a Gaussian process with zero mean and a non-degenerate covariance function.

R.2 Functions $\theta \mapsto \widehat{m}(\theta)$ and $\theta \mapsto m(\theta)$ admit continuous gradients $\nabla_\theta \widehat{m}(\theta)$ and $\nabla_\theta m(\theta)$ over the domain Θ, with probability one, where the former is a uniformly consistent estimate of the latter, namely uniformly in $\theta \in \Theta$

$$\nabla_\theta \widehat{m}(\theta) = \nabla_\theta m(\theta) + o_P(1).$$

Moreover, the norm of the gradient $\|\nabla_\theta m(\theta)|$ is bounded away from zero.

R.3 Weighting functions satisfy uniformly in $\theta \in \Theta$

$$s(\theta) = \sigma(\theta) + o_p(1), \quad w(\theta) = \omega(\theta) + o_p(1),$$

where $\sigma(\cdot) \geq 0$ and $\omega(\cdot) \geq 0$ are continuous functions bounded away from zero.

In Condition R.1, we require the estimates of the inequality-generating functions to satisfy a uniform central limit theorem. There are plenty of sufficient conditions for this to hold provided by the theory of empirical processes. In our example, this condition will follow from asymptotic normality of the estimates of the mean returns and covariance of returns. In Condition R.2, we require that gradient of the estimate of the inequality-generating function is consistent for the gradient of the inequality-generating function. Moreover, we require that the minimal eigenvalue of $\nabla_\theta m(\theta)\nabla_\theta m(\theta)'$ is bounded away from zero, which is an identification condition that allows us to estimate, at a usual speed, the boundary of the set Θ_0, which we define as

$$\partial\Theta_0 := \{\theta \in \Theta : m(\theta) = 0\}.$$

In Condition R.3, we require that the estimates of the weight functions are consistent for the weight functions, which are well-behaved.

Under these conditions we can state the following general result.

THEOREM 1 *(Limit Laws of LR and W Statistics). Under R.1-R.3*

$$\mathcal{L}_n =_d \mathcal{L} + o_p(1), \quad \mathcal{L} = \sup_{\theta \in \partial\Theta_0} \left[\frac{G(\theta)}{\sigma(\theta)}\right]_+^2, \tag{3.6}$$

$$\mathcal{W}_n =_d \mathcal{W} + o_p(1), \quad \mathcal{W} = \sup_{\theta \in \partial\Theta_0} \left[\frac{G(\theta)}{\|\nabla_\theta m(\theta)\| \cdot \omega(\theta)}\right]_+^2, \tag{3.7}$$

where both $\mathcal{W}$ and $\mathcal{L}$ have distribution functions that are continuous at their $(1-\alpha)$-quantiles for $\alpha < 1/2$. The two statistics are asymptotically equivalent under the following condition:

$$\mathcal{W}_n =_d \mathcal{L}_n + o_p(1) \quad \textit{if} \quad w(\theta) = \frac{\|\nabla_\theta m(\theta)\|}{\sigma(\theta)} \quad \textit{for each} \quad \theta \in \Theta.$$

We see from this theorem that the LR and W statistics converge in law to well-behaved random variables that are continuous transformations of the limit Gaussian process $G(\theta)$. Moreover, we see that under an appropriate choice of the weighting functions, the two statistics are asymptotically equivalent.

For our application to HJ and M sets, the following conditions will be sufficient

C.1 Estimator of the true parameter value γ_0 characterizing the inequality generating function $m(\theta) = m(\theta, \gamma_0)$, where γ_0 denotes the true parameter value, is such that $\sqrt{n}(\hat{\gamma} - \gamma_0) \to_d \Omega^{1/2}Z$, $Z = N(0, I_d)$.

C.2 Gradients $\nabla_\theta m(\theta, \gamma)$ and $\nabla_\gamma m(\theta, \gamma)$ are continuous over the compact parameter space $(\theta, \gamma) \in \Theta \times \Gamma$, where Γ is some set that includes an open neigborhood of γ_0. Moreover, the minimal eigenvalue of $\nabla_\theta m(\theta, \gamma)\nabla_\theta m(\theta, \gamma)'$ is bounded away from zero over $(\theta, \gamma) \in \Theta \times \Gamma$.

It is straightforward to verify that these conditions hold for the canonical versions of the HJ and M problems.

Under these conditions we immediately conclude that the following approximation is true uniformly in θ, that is, in the metric space of bounded functions $\ell^\infty(\Theta)$:

$$\begin{aligned} \sqrt{n}(\widehat{m}(\theta) - m(\theta)) &= \nabla_\gamma m(\theta, \bar{\gamma})'\sqrt{n}(\hat{\gamma} - \gamma_0) + o_p(1) & (3.8) \\ &= {}_d\nabla_\gamma m(\theta, \gamma_0)'\Omega^{1/2}Z + o_p(1), & (3.9) \end{aligned}$$

where $\nabla m(\theta, \bar{\gamma})$ denotes the gradient with each of its rows evaluated at a value $\bar{\gamma}$ on the line connecting $\hat{\gamma}$ and γ_0, where value $\bar{\gamma}$ may vary from row to row of the matrix. Therefore, the limit process in HJ and M examples takes the form:

$$G(\theta) = \nabla_\gamma m(\theta, \gamma_0)'\Omega^{1/2}Z. \tag{3.10}$$

This will lead us to conclude formally below that conclusions of Theorem 1 hold with

$$\begin{aligned} \mathcal{L} &= \sup_{\theta \in \partial\Theta_0} \left[\frac{\nabla_\gamma m(\theta, \gamma)'\Omega^{1/2}}{\sigma(\theta)} Z \right]_+^2, & (3.11) \\ \mathcal{W} &= \sup_{\theta \in \partial\Theta_0} \left[\frac{\nabla_\gamma m(\theta, \gamma)'\Omega^{1/2}}{\|\nabla_\theta m(\theta, \gamma)\| \cdot \omega(\theta)} Z \right]_+^2. & (3.12) \end{aligned}$$

A good strategy for choosing the weighting function for LR and W is to choose the studentizing Anderson-Darling weights

$$\begin{aligned}\sigma(\theta) &= \|\nabla_\gamma m(\theta,\gamma_0)'\Omega^{1/2}\|, && (3.13)\\ \omega(\theta) &= \frac{\|\nabla_\gamma m(\theta,\gamma_0)'\Omega^{1/2}\|}{\|\nabla_\theta m(\theta,\gamma_0)\|}. && (3.14)\end{aligned}$$

The natural estimates of these weighting functions are given by the following plug-in estimators:

$$\begin{aligned}s(\theta) &:= \|\nabla_\gamma m(\theta,\widehat{\gamma})'\widehat{\Omega}^{1/2}\|, && (3.15)\\ w(\theta) &:= \frac{\|\nabla_\gamma m(\theta,\widehat{\gamma})'\widehat{\Omega}^{1/2}\|}{\|\nabla_\theta m(\theta,\widehat{\gamma})\|}. && (3.16)\end{aligned}$$

We formalize the preceding discussion as the following corollary.

COROLLARY 1*(Limit Laws of LR and W statistics in HJ and M problems). Suppose that Conditions C.1-C.2 hold. Then conditions R.1 and R.2 hold with the limit Gaussian process stated in equation (3.10). Furthermore, the plug-in estimates of the weighting functions (3.15) and (3.16) are uniformly consistent for the weighting functions (3.13) and (3.14), so that Condition R.3 holds. Therefore, conclusions of Theorem 1 hold with the limit laws for our statistics given by the laws of random variables stated in equations (3.11) and (3.12).*

3.2.3 Basic Validity of the Confidence Regions

In this section we shall suppose that we have suitable estimates of the quantiles of LR and W statistics and will verify basic validity of our confidence regions. In the next section we will provide a construction of such suitable estimates by the means of bootstrap and simulation.

Our result is as follows.

THEOREM 2 *(Basic Inferential Validity of Confidence Regions). Suppose that for*

$\alpha < 1/2$ we have consistent estimates of quantiles of limit statistics $\mathcal{W}$ and $\mathcal{L}$, namely,

$$\hat{k}(1-\alpha) = k(1-\alpha) + o_p(1), \tag{3.17}$$

where $k(1-\alpha)$ is $(1-\alpha)$-quantile of either $\mathcal{W}$ or $\mathcal{L}$. Then as the sample size n grows to infinity, confidence regions R_{LR} and R_W cover Θ_0 with probability approaching $1-\alpha$:

$$\begin{aligned} Pr_P[\Theta_0 \subseteq R_{LR}] &= Pr_P[\mathcal{L}_n \leq \hat{k}(1-\alpha)] \rightarrow Pr_P[\mathcal{L} \leq k(1-\alpha)] = (1-\alpha)(3.18) \\ Pr_P[\Theta_0 \subseteq R_W] &= Pr_P[\mathcal{W}_n \leq \hat{k}(1-\alpha)] \rightarrow Pr_P[\mathcal{W} \leq k(1-\alpha)] = (1-\alpha)(3.19) \end{aligned}$$

The result further applies to HJ and M problems.

COROLLARY 2*(Limit Laws of LR and W statistics in HJ and M problems). Suppose that Conditions C.1-C.2 hold and that consistent estimates of quantiles of statistics (3.11) and (3.12) are available. Then conclusions of Theorem 2 apply.*

3.2.4 Estimation of Quantiles of LR and W Statistics by Bootstrap and Other Methods

In this section we show how to estimate quantiles of LR and W statistics using bootstrap, simulation, and other resampling schemes under general conditions. The basic idea is as follows: First, let us take any procedure that consistently estimates the law of our basic Gaussian process G or a weighted version of this process appearing in the limit expressions. Second, then we can show with some work that we can get consistent estimates of the laws of LR and W statistics, and thus also obtain consistent estimates of their quantiles. It is well-known that there are many procedures for accomplishing the first step, including such common schemes as the bootstrap, simulation, and subsampling, including both cross-section and time series versions.

In what follows, we will ease the notation by writing our limit statistics as a special case of the following statistic:

$$\mathcal{S} = \sup_{\theta \in \partial\Theta_0} [V(\theta)]_+, \quad V(\theta) = \tau(\theta)G(\theta). \tag{3.20}$$

Thus, $\mathcal{S} = \mathcal{L}$ for $\tau(\theta) = 1/s(\theta)$ and $\mathcal{S} = \mathcal{W}$ for $\tau(\theta) = 1/[\|\nabla_\theta m(\theta)\| \cdot \omega(\theta)]$. We take τ to be a continuous function bounded away from zero on the parameter space. We also need to introduce the following notations and concepts. Our process V is a random element that takes values in the metric space of continuous functions $C(\Theta)$ equipped with the uniform metric. The underlying measure space is $(\Omega, \mathcal{F})$ and we denote the law of V under the probability measure P by the symbol $\mathcal{Q}_V$.

Suppose we have an estimate $\mathcal{Q}_{V^*}$ of the law $\mathcal{Q}_V$ of the Gaussian process V. This estimate $\mathcal{Q}_{V^*}$ is a probability measure generated as follows. Let us fix another measure space $(\Omega', \mathcal{F}')$ and a probability measure P^* on this space, then given a random element V^* on this space taking values in $C(\Theta)$, we denote its law under P^* by $\mathcal{Q}_{V^*}$. We thus identify the probability measure P^* with a data-generating process by which we generate draws or realizations of V^*. This identification allows us to encompass such methods of producing realizations of V^* as the bootstrap, subsampling, or other simulation methods. We require that the estimate $\mathcal{Q}_{V^*}$ is consistent for $\mathcal{Q}_V$ in any metric ρ_K metrizing weak convergence, where we can take the metric to be the Kantarovich-Rubinstein metric. Let us mention right away that there are many results that verify this basic consistency condition for various rich forms of processes V and various bootstrap, simulation, and subsampling schemes for estimating the laws of these processes, as we will discuss in more detail below.

In order to recall the definition of the Kantarovich-Rubinstein metric, let $\theta \mapsto v(\theta)$ be an element of a metric space (M, d), and $Lip(M)$ be a class of Lipschitz functions $\varphi : M \to \mathbb{R}$ that satisfy:

$$|\varphi(v) - \varphi(v')| \leq d(v, v') \wedge 1, \quad |\varphi(v)| \leq 1,$$

The Kantarovich-Rubinstein distance between probability laws $\mathcal{Q}$ and $\mathcal{Q}'$ is

$$\rho_K(\mathcal{Q}, \mathcal{Q}'; M) := \sup_{\varphi \in Lip(M)} |E_{\mathcal{Q}}\varphi - E_{\mathcal{Q}'}\varphi|.$$

As stated earlier, we require that the estimate $\mathcal{Q}_{V^*}$ is consistent for $\mathcal{Q}_V$ in the metric ρ_K, that is

$$\rho_K(\mathcal{Q}_{V^*}, \mathcal{Q}_V; C(\Theta)) = o_p(1). \tag{3.21}$$

Let $\mathcal{Q}_{\mathcal{S}}$ denote the probability law of $\mathcal{S} = \mathcal{W}$ or $\mathcal{L}$, which is in turn induced by the law $\mathcal{Q}_V$ of the Gaussian process V. We need to define the estimate $\mathcal{Q}_{\mathcal{S}^*}$ of this law. First, we define the following plug-in estimate of the boundary set $\partial\Theta_0$, which we need to state here:

$$\widehat{\partial\Theta_0} = \{\theta \in \Theta : \widehat{m}(\theta) = 0\}. \tag{3.22}$$

This estimate turns out to be consistent at the usual root-n rate, by the argument like the one given in Chernozhukov, Hong, and Tamer (2007). Then define $\mathcal{Q}_{\mathcal{S}^*}$ as the law of the following random variable

$$\mathcal{S}^* = \sup_{\theta \in \widehat{\partial\Theta_0}} [V^*(\theta)]_+ \tag{3.23}$$

In this definition, we hold the hatted quantities fixed, and the only random element is V^* that is drawn according to the law $\mathcal{Q}_{V^*}$.

We will show that the estimated law $\mathcal{Q}_{\mathcal{S}^*}$ is consistent for $\mathcal{Q}_{\mathcal{S}}$ in the sense that

$$\rho_K(\mathcal{Q}_{\mathcal{S}^*}. \mathcal{Q}_{\mathcal{S}}; \mathbb{R}) = o_p(1). \tag{3.24}$$

Consistency in the Kantarovich-Rubinstein metric in turn implies consistency of the estimates of the distribution function at continuity points, which in turn implies consistency of the estimates of the quantile function.

Equipped with the notations introduced above we can now state our result.

THEOREM 3 *(Consistent Estimation of Quantiles) Suppose Conditions R.1-R.3* *hold, and any mechanism, such as bootstrap or other method, is available, which provides a consistent estimate of the law of our limit Gaussian processes V , namely equation (3.21) holds. Then, the estimates of the laws of the limit statistics $\mathcal{S} = \mathcal{W}$ or $\mathcal{L}$ defined above are consistent in the sense of equation (3.24). As a consequence, we have that the estimates of the quantiles are consistent in the sense of equation (3.17).*

We now specialize this result to the HJ and M problems. We begin by recalling that our estimator satisfies

$$\sqrt{n}(\hat{\gamma} - \gamma) =_d \Omega^{1/2} Z + o_p(1).$$

Then our limit statistics take the form:

$$\mathcal{S} = \sup_{\theta \in \partial \Theta_0} [V(\theta)]_+^2, \quad V(\theta) = t(\theta)' Z,$$

where $t(\theta)$ is a vector valued weight function, in particular, for $\mathcal{S} = \mathcal{L}$ we have $t(\theta) = (\nabla_\gamma m(\theta, \gamma)' \Omega^{1/2})/\sigma(\theta)$ and for $\mathcal{S} = \mathcal{W}$ we have $t(\theta) = (\nabla_\gamma m(\theta, \gamma)' \Omega^{1/2})/(\|\nabla_\theta m(\theta, \gamma)\| \cdot \omega(\theta))$. Here we shall assume that we have a consistent estimate $\mathcal{Q}_{Z^*}$ of the law $\mathcal{Q}_Z$ of Z, in the sense that,

$$\rho_K(\mathcal{Q}_{Z^*}, \mathcal{Q}_Z) = o_p(1). \tag{3.25}$$

There are many methods that provide such consistent estimates of the laws. Bootstrap is known to be valid for various estimation methods (van der Vaart and Wellner (1996)); simulation method that simply draws $Z \sim N(0, I)$ is another valid method; and subsampling is another rather general method (Politis and Romano (1994)). Next, the estimate $\mathcal{Q}_{V^*}$ of the law $\mathcal{Q}_{V^*}$ is then defined as:

$$V^*(\theta) = \widehat{t}(\theta)' Z^*, \tag{3.26}$$

where $\widehat{t}(\theta)$ is a vector valued weighting function that is uniformly consistent for the weighting function $t(\theta)$. In this definition we hold the hatted quantity fixed, and the only random element is Z^* that is drawn according to the law $\mathcal{Q}_{Z^*}$. Then, we define the random variable

$$\mathcal{S}^* = \sup_{\theta \in \widehat{\partial\Theta_0}} [V^*(\theta)]_+^2,$$

and use its law $\mathcal{Q}_{\mathcal{S}^*}$ to estimate the law $\mathcal{Q}_{\mathcal{S}}$.

We can now state the following corollary.

Corollary 3 *(Consistent Estimation of Quantiles in HJ and M problems) Suppose Conditions C.1-C.2 hold, and any mechanism, such as bootstrap or other method, that provides a consistent estimate of the law of Z is available, namely equation (3.25) holds. Then, this provides us with a consistent estimate of the law of our limit Gaussian process G, namely equation (3.21) holds. Then, all of the conclusions of Theorem 3 hold.*

3.3 Empirical Example

As an empirical example we use HJ bounds which are widely used in testing asset pricing models. In order to keep results comparable, the sample used in this section is very similar to data used in Hansen and Jagannathan (1991). The two asset series used are annual treasury bond returns and annual NYSE value-weighted dividend included returns. These nominal returns are converted to real returns by using implicit price deflator based on personal consumption expenditures as in Hansen and Jagannathan (1991). Asset returns are from CRSP, and the implicit price deflator is available from St. Louis Fed and based on National Income and Product Accounts of United States. We use data for the time period 1959-2006 (inclusive).

Figure 3-1 simply traces out the mean-standard deviation pairs which satisfy

$$m(\theta, \hat{\gamma}) = 0$$

where $\hat{\gamma}$ is estimated using sample moments.

Figure 3-2 represents the uncertainity caused by the estimation of γ. To estimate the distribution of $\hat{\gamma}$ bootstrap method is used. Observations are drawn with replacement from the bivariate time series of stock and bond returns. 100 bootstraps result in 100 $\hat{\gamma}$. The resulting HJ bounds are included in the figure.

In Figure 3-3 in addition to the bootstrapped curves 90% confidence region based on LR statistic is presented. LR based confidence region covers most of the bootstrap draws below the HJ bounds as expected. An attractive outcome of using this method is that the resulting region does not include any unnecessary areas that is not covered by bootstrap draws.

Figure 3-4 plots 90% confidence region based on unweighted LR statistic. Comparison of Figure 3-3 and Figure 3-4 reveals that precision weighting plays a very important role in delivering good confidence sets. Without precision weighting LR statistic delivers a confidence region that includes unlikely regions in the parameter space where standard deviation of the discount factor is zero. On the other hand precision weighted LR based confidence region is invariant to parameter transformations, for example, changes in units of measurement. This invariance to parameter transformations is the key property of a statistic to deliver desirable confidence regions that does not cover unnecessary areas.

Figure 3-5 plots confidence region based on Wald-based statistic with no precision weighting. This is identical to the confidence region based on Hausdorff distance. Similar to Figure 3-4 this region covers a large area of the parameter space where no bootstrap draws appear. This picture reveals a key weakness of using an unweighted Wald-based statistic or Hausdorff distance to construct confidence regions. These methods are not invariant to parameter transformations which results in confidence regions with undesirable qualities that cover unnecessary areas in the parameter space. The problem in Figure 3-4 and Figure 3-5 are of similar nature. In both of these cases the statistics underlying the confidence regions are not invariant to parameter transformations therefore when drawing confidence regions uncertainity in one part of the plot is assumed to be identical to uncertainity in other parts of the plot. However a quick look at the Figure 3-2 reveals that uncertainity regarding the location of the HJ bound varies for a given mean or standard deviation of the stochastic discount factor.

Figure 3-6 plots the confidence region based on weighted Wald statistic. Weighting fixes the problem and generates a statistic that is invariant to parameter transformations. The resulting confidence set looks very similar to weighted LR based confidence set in Figure 3-3 as it covers most of the bootstrap draws below the HJ bounds and does not include unnecessary regions in the parameter space.

3.4 Conclusion

In this paper we provided various inferential procedures for inference on sets that solve a system of inequalities. These procedures are useful for inference on Hansen-Jagannathan mean-variance sets of admissible stochastic discount factors and Markowitz mean-variance sets of admissible portfolios.

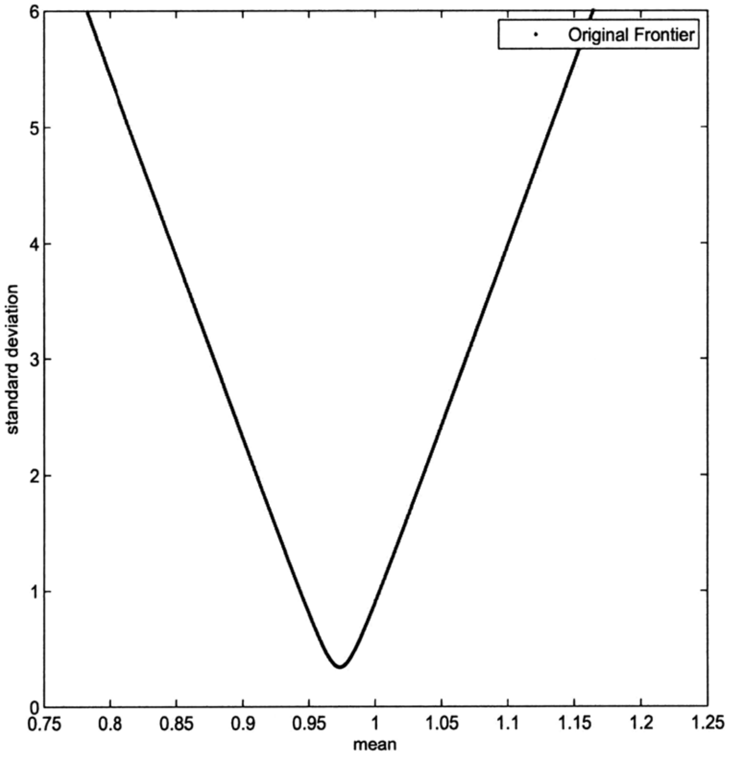


Figure 3-1: Estimated HJ Bounds

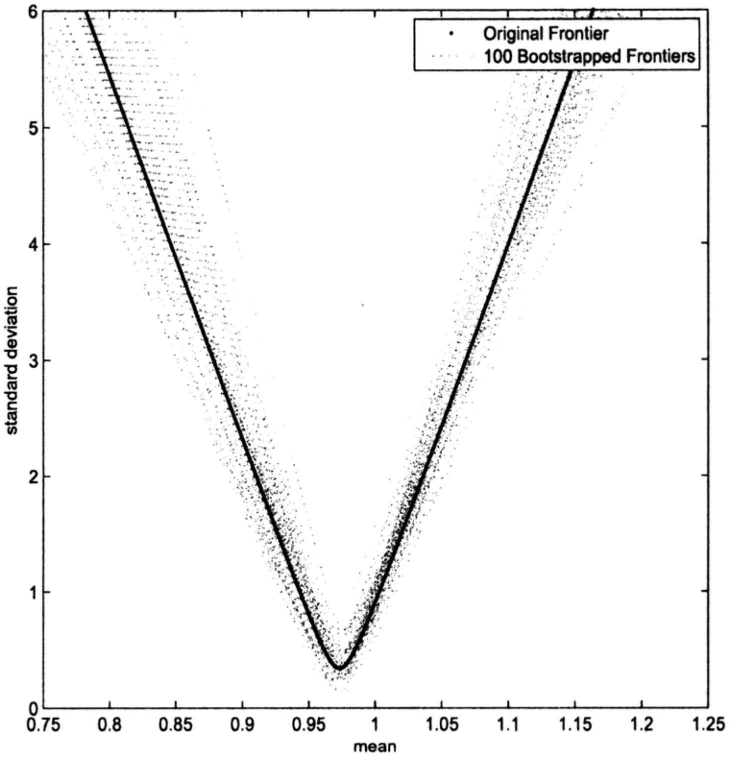


Figure 3-2: Estimated HJ Bounds and Bootstrap Draws

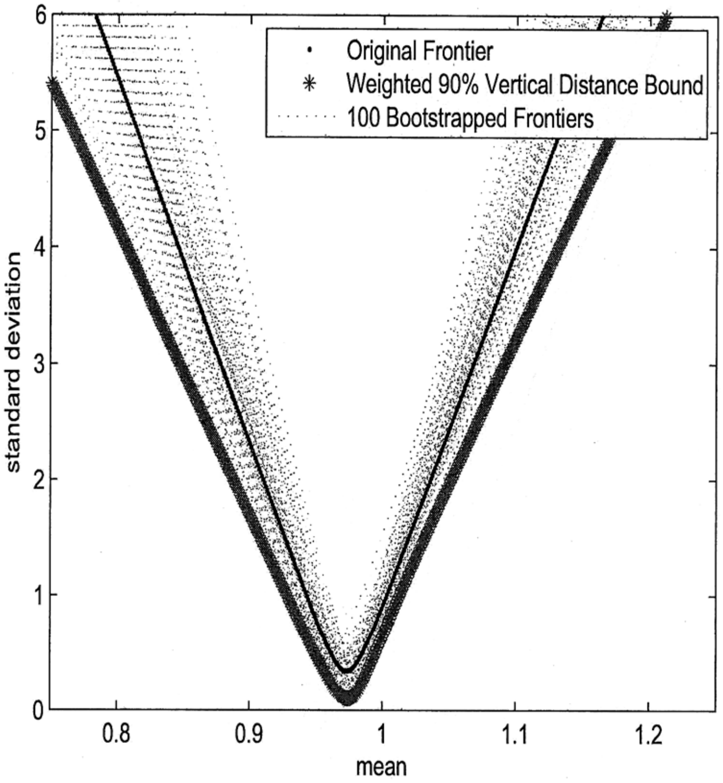


Figure 3-3: 90% Confidence Region using LR Statistic

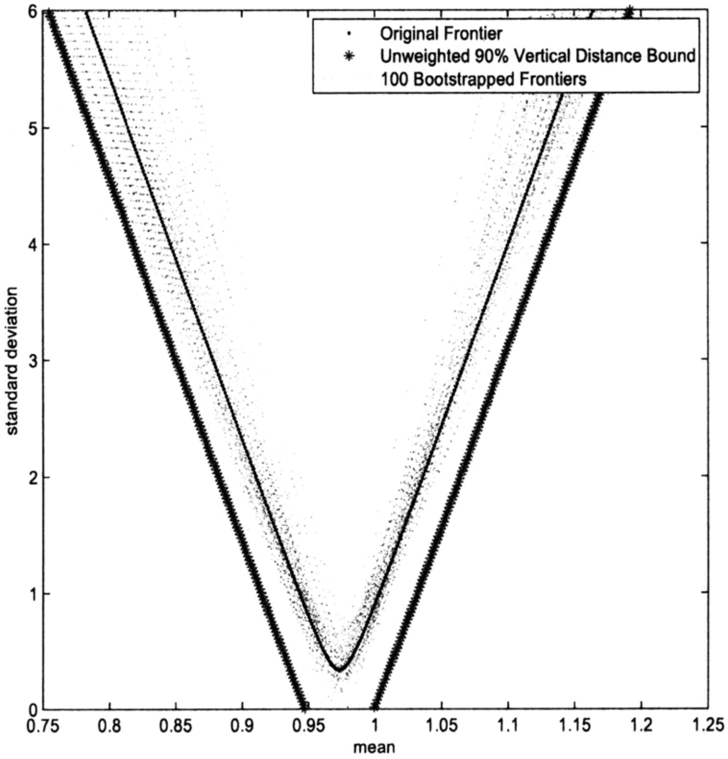


Figure 3-4: 90% Confidence Region using Unweighted LR Statistic

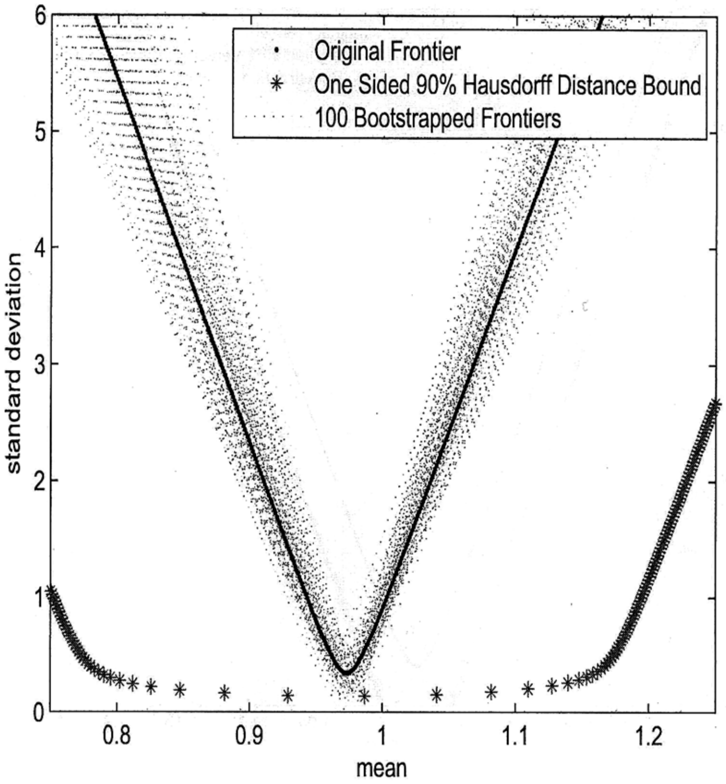


Figure 3-5: 90% Confidence Region using Unweighted W Statistic (H-Distance)

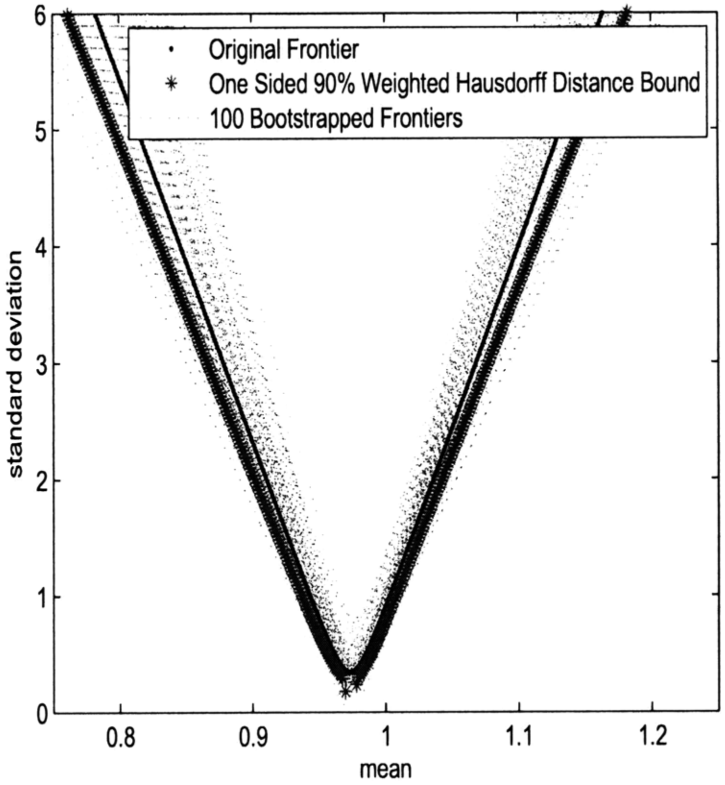


Figure 3-6: 90% Confidence Region using Weighted W Statistic

Appendix A

Proofs for Chapter 1

Derivation of the Rate in Example 5:

Denote by $\tilde{h}^m(z,\theta,P)$ the projection of $h(z,\theta,P)$ onto the space of B-splines with nonnegative coefficients with basis functions $\psi^m(z)$. Suppose there is $\theta_0 \in \Theta_I \backslash \Theta_{I,n}$, so that for all $z \in \mathcal{Z}$ $h(z,\theta_0,P_n) \geq 0$, but $\tilde{h}^{m_n}(\bar{z},\theta_0,P_n) < 0$ for some $\bar{z} \in \mathcal{Z}$. By uniform approximation through splines $|\tilde{h}^{m_n}(\bar{z},\theta_0,P_n) - h(\bar{z},\theta_0,P_n)| = O(c_m)$. Let θ_1 be the point closest to θ_0 with respect to the Euclidean metric for which $\tilde{h}^{m_n}(\bar{z},\theta_0,P_n) \geq 0$. By standard arguments,

$$(\theta_0 - \theta_1) \approx D_n(z,\theta_0)(D_n(z,\theta_0)'D_n(z,\theta_0))^{-1}[\tilde{h}^{m_n}(\bar{z},\theta_0,P_n) - h(\bar{z},\theta_0,P_n)]$$

so that by our assumptions on the rate of D_n, $\theta_0 - \theta_1 = O\left(\frac{c_m}{a_n^{1/r}}\right)$. For $\theta_A \in \Theta_{I,n} \backslash \Theta_I$, the analogous argument goes through with only slight modifications since eventually $\kappa < \tau_m$, so that taking both steps together, we can establish the rate stated above □

Proof of Theorem 1:

For part (i), we will show that with probability approaching 1, $\Theta_{I,n} \subset \hat{\mathcal{C}}_n$ and $\hat{\mathcal{C}}_n \subset \Theta_{I,n}^K$ for any $K > 0$, where $\Theta_{I,n}^K$ denotes the K blow-up of $\Theta_{I,n}$ with respect to the renormalized Hausdorff distance $\Theta_{I,n}^K := \{\theta \in \Theta : \mu_n^{-1/r}\varrho_n(\theta,\Theta_{I,n}) \leq K\}$.

First note that by Condition 5, $\hat{c}_n \geq \sup_{\theta\in\Theta_{I,n}} n\hat{Q}_n(\theta)$ with probability converging to one, so that $\Theta_{I,n} \subset \hat{\mathcal{C}}_n$, implying $d(S_{n\theta}\theta, S_{n\theta}\hat{\mathcal{C}}_n) = 0$ for all $\theta \in \Theta_{I,n}$ with probability approaching 1.

In order to prove that the set estimator approaches $\Theta_{I,n}$ from the outside, we have to show that for any $K > 0$, $\hat{\mathcal{C}}_n \subset \Theta_{I,n}^K$ with probability approaching 1. By construction, $\mu_n^{-1} \sup_{\theta\in\hat{\mathcal{C}}_n} n\hat{Q}_n(\theta) = \frac{\hat{c}_n}{\mu_n}$ which is $o_p(1)$ by assumption 5. On the other hand, by uniform convergence from Condition 3 (c), for any choice of $\eta > 0$ and n large enough,

$$\mu_n^{-1} \inf_{\theta\in\Theta\backslash\Theta_{I,n}^K} n\hat{Q}_n(\theta) \geq \inf_{\theta\in\Theta\backslash\Theta_{I,n}^K} (\gamma_n(\theta)+\alpha_n\delta_n(\theta))(1-\eta)-\eta \geq \inf_{\theta\in\Theta\backslash\Theta_{I,n}^K} \gamma(\theta)(1-3\eta)-3\eta$$

where the last step uses Condition 3 (b) and $\alpha_n \to 0$ and $\delta_n(\theta)$ is uniformly bounded on $\tilde{\Theta}_n^\delta$.

By epi-convergence from Condition 3 (b), $\arg\inf_{\theta\in\Theta} \gamma(\theta) = \Theta_I$, so that $\gamma(\theta) > 0$ for any $\theta \in \Theta\backslash\Theta_I$. Since $\Theta_I \subset \Theta_{I,n}$, and by compactness of $\Theta\backslash\Theta_{I,n}^K$, $\inf_{\theta\in\Theta\backslash\Theta_{I,n}^K} \gamma(\theta) > \varepsilon$ for some $\varepsilon > 0$. Hence choosing e.g. $\eta = \frac{\varepsilon}{4(1+\varepsilon)}$, we have that with probability converging to 1, $\hat{\mathcal{C}}_n \subset \Theta_{I,n}^K$, implying $d(S_{n\theta}\theta, S_{n\theta}\Theta_{I,n}^K) = 0$ for all $\theta \in \hat{\mathcal{C}}_n$. Since K was arbitrary, this establishes part (i).

We will now prove part (ii). By the same argument as above, $\Theta_{I,n} \subset \hat{\mathcal{C}}_n$ with probability converging to 1. Now, let $K_n = \left(\frac{\hat{c}_n}{\kappa_1\mu_n}\right)^r$ which converges to zero in probability by Condition 5 so that for any $\varepsilon > 0$ and n large enough, $P(K_n \geq \delta) < \varepsilon$. Then by Condition 4,

$$\inf_{\theta\in\Theta\backslash\Theta_{I,n}^{K_n}} n\hat{Q}_n(\theta) \geq \kappa_1\mu_n(K_n \wedge \delta)^{1/r} = \hat{c}_n$$

with probability approaching 1. On the other hand, by definition of the set estimator, $\sup_{\theta\in\hat{\mathcal{C}}_n} n\hat{Q}_n(\theta) = \hat{c}_n$ so that $P(\hat{\mathcal{C}}_n \subset \Theta_{I,n}^{K_n}) \to 1$. Hence, $\mu_n^{-1/r}\varrho_n(\hat{\mathcal{C}}_n, \Theta_{I,n}) = O_P\left(\frac{\hat{c}_n}{\mu_n}\right)^{1/r}$, which completes the proof □

Proof of Proposition 1:

In order to show uniform convergence for the criterion function, we will first prove the following Lemmas:

Lemma 4 *The orthogonal projection of $\sqrt{n}\bar{g}+\zeta$ onto a cone $\mathcal{C}$ with respect to the scalar product $\langle\cdot,\cdot\rangle_W$ is a contraction, i.e. for any $\zeta,\tilde{\zeta}$,*

$$\|\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)-\Pi(\sqrt{n}\bar{g}+\tilde{\zeta}|\mathcal{C},W)\|_W^2\leq\|\zeta-\tilde{\zeta}\|_W^2$$

Proof: By an orthogonal projection result for convex cones (e.g. Lemma 2.7.5 in Stoer and Witzgall (1970)), we can write for any ζ

$$\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)=\zeta-\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C}^\circ,W)$$

where $\mathcal{C}^\circ$ denotes the polar cone to $\mathcal{C}$ with respect to the scalar product $\langle\cdot,\cdot\rangle_W$ induced by W, and

$$\langle\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W).\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C}^\circ,W)\rangle_W=0 \tag{A.1}$$

Now, instead of calculating the moments of ζ and $\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)$, we will look at the differences between two independent draws $\zeta,\tilde{\zeta}$ from the same distribution, which have mean zero by construction. For any pair $\zeta,\tilde{\zeta}$ we have

$$\begin{aligned}
\|\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W) &- \Pi(\sqrt{n}\bar{g}+\tilde{\zeta}|\mathcal{C},W)\|_W^2=\|\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C}^\circ,W)-\Pi(\sqrt{n}\bar{g}+\tilde{\zeta}|\mathcal{C}^\circ,W)\|_W^2\\
&\quad+\|\zeta-\tilde{\zeta}\|_W^2-2(\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C}^\circ,W)-\Pi(\sqrt{n}\bar{g}+\tilde{\zeta}|\mathcal{C}^\circ,W))'W(\zeta-\tilde{\zeta})\\
&= \|\zeta-\tilde{\zeta}\|_W^2+\|\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C}^\circ,W)-\Pi(\sqrt{n}\bar{g}+\tilde{\zeta}|\mathcal{C}^\circ,W)\|_W^2\\
&\quad+2\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C}^\circ,W)'W\Pi(\sqrt{n}\bar{g}+\tilde{\zeta}|\mathcal{C},W)\\
&\quad+2\Pi(\sqrt{n}\bar{g}+\tilde{\zeta}|\mathcal{C}^\circ,W)'W\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)\\
&\leq \|\zeta-\tilde{\zeta}\|_W^2+\|\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C}^\circ,W)-\Pi(\sqrt{n}\bar{g}+\tilde{\zeta}|\mathcal{C}^\circ,W)\|_W^2\\
&\leq \|\zeta-\tilde{\zeta}\|_W^2
\end{aligned}$$

where the second equality holds by A.1, and the first inequality uses that by definition, any vector in $\mathcal{C}$ forms an obtuse angle with respect to $\langle\cdot,\cdot\rangle_W$ with any vector in the corresponding polar cone $\mathcal{C}^\circ$ $\square$

Lemma 5 *We can bound the expectation of the norm of the projection by*

$$\mathbb{E}\|\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)-\mathbb{E}[\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)]\|^2\leq\mathbb{E}\|\zeta\|_W^2$$

PROOF: Noting that for two independent draws X_1, X_2 from the same distribution, $\mathbb{E}\|X_1-X_2\|^2=2\mathrm{Var}(X_1)$, by Lemma 4,

$$\begin{aligned}\mathbb{E}\|\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)-\mathbb{E}[\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)]\|^2 &= \frac{1}{2}\mathbb{E}\|\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)-\Pi(\sqrt{n}\bar{g}+\tilde{\zeta}|\mathcal{C},W)\|^2\\ &= \frac{1}{2}\mathbb{E}\|\zeta-\tilde{\zeta}\|_W^2=\mathbb{E}\|\zeta\|_W^2\end{aligned}$$

since the expectation of ζ equals zero $\square$

Lemma 6 *If the fourth moments of $\zeta_n(\theta)$ are bounded uniformly in θ, then the fourth moments of $\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)$ are also uniformly bounded.*

PROOF: As in the preceding Lemma, and noting that for two i.i.d. draws X_1, X_2, $\mathbb{E}\|X_1-X_2\|^4=2\mathbb{E}\|X_1-\mathbb{E}[X_1]\|^4+2\left(\mathbb{E}\|X_1-\mathbb{E}[X_1]\|^2\right)^2$. we can produce a very generous (but finite) bound using Lemma 4 for independent draws $\zeta,\tilde{\zeta}$:

$$\begin{aligned}\mathbb{E}\|\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)-\mathbb{E}[\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)]\|^4 &= \frac{1}{2}\mathbb{E}\|\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)-\Pi(\sqrt{n}\bar{g}+\tilde{\zeta}|\mathcal{C},W)\|^4\\ &\quad-\left(\mathbb{E}\|\Pi(\sqrt{n}\bar{g}+\zeta|\mathcal{C},W)\|\right)^2\\ &\leq \frac{1}{2}\mathbb{E}\|\zeta-\tilde{\zeta}\|_W^4\leq\mathbb{E}\|\zeta\|_W^4+\left(\mathbb{E}\|\zeta\|_W^2\right)^2\end{aligned}$$

where both terms on the right-hand side of the last inequality were assumed to be finite $\square$

Lemma 7 *Under Condition 3,*

$$\mathrm{Var}\left(W_n(\theta)^{1/2}\sqrt{n}\hat{t}_n(\theta)\right) \leq \mathrm{Var}\left(W_n(\theta)^{1/2}\zeta_n(\theta)\right)$$

in the positive definite matrix sense.

PROOF: Note that $\hat{t}_n$ is the projection of $\sqrt{n}\bar{g}_n(\theta) + \zeta_n(\theta)$ onto a polyhedral cone. Hence, for each value of the vector and the corresponding set of constraints $\mathcal{J}_n \subset \{1, \dots, m_n\}$ defining the face of the cone $\sqrt{n}\sqrt{n}\bar{g}_n(\theta) + \zeta_n(\theta)$ is projected onto, $\sqrt{n}\hat{t}_n$ constitutes an orthogonal projection with respect to the distance weighted by $W_n(\theta)$ onto the linear subspace $\mathcal{L} = \mathrm{span}(\{e_j : j \in \mathcal{J}\})$, where e_j denotes the jth unit vector. For projections onto linear spaces, it is known (see e.g. Malinvaud (1980), section 6.4) that $\mathrm{Var}(\sqrt{n}\bar{g}_n(\theta) + \zeta_n(\theta)|\mathcal{J}_n) - \mathrm{Var}(\sqrt{n}\hat{t}_n|\mathcal{J}_n)$ is positive definite. By a similar argument, $\mathrm{Var}(\mathbb{E}[\zeta_n(\theta)|\mathcal{J}_n]) \geq \mathrm{Var}(\mathbb{E}[\hat{t}_n|\mathcal{J}_n])$, so that the desired conclusion follows from the conditional variance identity $\mathcal{J}_n$ □

Lemma 8 *Define*

$$T_{1n}(\theta) := \frac{\sqrt{n}}{\mu_n(\theta)}\langle\bar{g}_n(\theta) - t_{0n}(\theta), \zeta_n(\theta)\rangle_W$$

and

$$T_{2n}(\theta) := \frac{\sqrt{n}}{\mu_n(\theta)}\langle\bar{g}_n(\theta) - t_{0n}(\theta). \hat{t}_n(\theta) - \mathbb{E}[\hat{t}_n(\theta)]\rangle_W$$

Then, under the conditions of Proposition 1,

$$\sup_{\theta\in\Theta} h(\theta)^{-1}|T_{1n}(\theta)| \xrightarrow{p} 0 \text{ and } \sup_{\theta\in\Theta} h(\theta)^{-1}|T_{2n}(\theta)| \xrightarrow{p} 0.$$

PROOF: We will first show pointwise convergence, and then show that the sequence is asymptotically tight, so that uniformity follows e.g. from Theorem 7.1 in Billingsley (1999). For the argument based on Prohorov's Theorem we can in fact dispense of measurability conditions via the Hoffmann-Jørgensen approach using convergence in outer measure, see e.g. van der Vaart and Wellner (1996).

By inspection, T_{1n} has mean zero, and we can bound the variance by

$$\begin{aligned} \mathbb{E}[T_1^2] &= \mu_n(\theta)^{-2}\mathbb{E}[\left|\langle\sqrt{n}(\bar{g}_n(\theta) - t_{0n}), \zeta_n(\theta)\rangle_W\right|^2] \\ &\leq n\mu_n(\theta)^{-2}(\bar{g}_n(\theta) - t_{0n})'W_n(\theta)^{1/2}\mathbb{E}[W_n(\theta)^{1/2}\zeta_n(\theta)\zeta_n(\theta)'W_n(\theta)^{1/2}]W_n(\theta)^{1/2}(\bar{g}_n(\theta) - t_{0n}) \\ &\preceq n\mu_n(\theta)^{-2}(\bar{g}_n(\theta) - t_{0n})'W_n(\theta)(\bar{g}_n(\theta) - t_{0n}) = \mu_n(\theta)^{-1}\gamma_n(\theta) \end{aligned}$$

where "$\preceq$" means "less than or equal up to a multiplicative constant," and since by Condition 3 (b), the maximal eigenvalue of $\mathbb{E}[W_n(\theta)^{1/2}\zeta_n(\theta)\zeta_n(\theta)'W_n(\theta)^{1/2}]$ is uniformly bounded. Now note that by construction $h_n(\theta)^{-1}\gamma_n(\theta)$ is uniformly bounded over $\theta \in \Theta$. Since $\mu_n \to \infty$ by Assumption 5, the variance of $h_n(\theta)^{-1}T_{1n}(\theta)$ goes to zero for all $\theta \in \Theta$ so that, by Chebyshev's inequality, for all $\theta \in \Theta$ $h_n(\theta)^{-1}T_{1n}(\theta) \xrightarrow{p} 0$. Using the same argument and Lemma 7, we also have pointwise convergence for $h_n(\theta)^{-1}T_{2n}(\theta)$.

In order to prove tightness, note that we can use the Cauchy-Schwarz Inequality to bound

$$\begin{aligned} |T_{1n}(\theta)| &= \sqrt{T_{1n}(\theta)^2} = \frac{1}{\mu_n(\theta)}\sqrt{\left|\langle\sqrt{n}(\bar{g}_n(\theta) - t_{0n}(\theta)), \zeta_n(\theta)\rangle_W\right|^2} \\ &\leq \frac{\sqrt{nm_n}}{\mu_n(\theta)}\|\bar{g}_n(\theta) - t_{0n}(\theta)\|_W m_n^{-1/2}\,\|\zeta_n(\theta)\|_W \leq \alpha_n^{1/2}\gamma_n(\theta)^{1/2}m_n^{-1/2}\,\|\zeta_n(\theta)\|_W \end{aligned}$$

Since $h_n(\theta)^{-1}\gamma_n(\theta)$ is uniformly bounded in Θ, it suffices show that $\frac{1}{\sqrt{m_n}}\,\|\zeta_n(\theta)\|_W$ is tight, which follows from Condition 3, part(d) by

$$m_n^{-1}\|\zeta_n(\theta)\|_W \leq C_W^* \max_{m\leq m_n} \zeta_{mn}(\theta)^2 \leq \left(\max_{m\leq m_n} |\zeta_{mn}(\theta)|\right)^2$$

where $C_W^* := \max\text{eig}(W)$ is the largest eigenvalue of W. Therefore, $h_n(\theta)^{-1}T_1$ is concentrated on a compact. Using Lemma 5, we get the analogous result for T_2 □

Lemma 9 *Under the conditions of Proposition 1,*

$$T_{3n} := \mu_n^{-1}\left(\|\zeta_n - \sqrt{n}(\hat{t}_n - t_{0n})\|_W^2 - \mathbb{E}\left[\|\zeta_n - \sqrt{n}(\hat{t}_n - t_{0n})\|_W^2\right]\right) \xrightarrow{p} 0$$

uniformly in $\theta \in \tilde{\Theta}_n^\delta$.

Proof: Denote $\Delta := \Pi(\bar{g} + \zeta|\mathcal{C}, W) - \mathbb{E}[\Pi(\bar{g} + \zeta|\mathcal{C}, W)]$ By Lemma 6 and Assumption 3 (b), the fourth moment of Δ is bounded, and hence by the triangle inequality and Chebyshev's Inequality, $m_n^{-1}(\|\Delta\|_W^2 - \mathbb{E}\|\Delta\|_W^2) \xrightarrow{p} 0$. Notice also that by the triangle inequality

$$m_n^{-1}|\|\Delta\|_W^2 - \mathbb{E}\|\Delta\|_W^2| \leq m_n^{-1}\|\Delta\|_W^2 + m_n^{-1}\mathbb{E}\|\Delta\|_W^2$$

where $m_n^{-1}\|\Delta\|_W^2$ is tight by the same argument as in the previous lemma, and $m_n^{-1}\mathbb{E}\|\Delta\|_W^2$ is uniformly bounded by Condition 3 (b). Since $\frac{m_n}{\mu_n} = \alpha_n \to 0$, $T_{3n} = \alpha_n m_n^{-1}(\|\Delta\|_W^2 - \mathbb{E}\|\Delta\|_W^2)$ converges to zero in probability uniformly in θ □

Proof of Proposition 1: The proof of uniform convergence is similar to that for GMM in Han and Phillips (2006), except that we have to bound terms using the contraction argument at several steps since there is no closed-form expression for projections of the noise part. By Assumption 4 and Assumption 3 (d),

$$n\|\hat{g}_n - t^*(\theta)\|_{W_n}^2 - n\|\hat{g}_n(\theta) - t^*(\theta)\|_W^2 \xrightarrow{p} 0$$

uniformly in θ, so that in the following we can hold $W_n(\theta)$ fixed at $W(\theta)$. I will also suppress subscripts and arguments wherever possible with the understanding that all quantities are evaluated at sample size n and the parameter $\theta \in \tilde{\Theta}_n^\delta$. We can then rewrite

$$\begin{aligned} nQ_n &= \|\sqrt{n}(\bar{g} - t^*) + \zeta\|_W^2 \\ &= \|\sqrt{n}(\bar{g} - t_0) + (\sqrt{n}(\bar{g} - t^*) + \zeta) - \sqrt{n}(\bar{g} - t_0)\|_W^2 \\ &= n\|\bar{g} - t_0\|_W^2 + \|\zeta - \sqrt{n}(\nu^* - t_0)\|_W^2 + 2\sqrt{n}\langle \bar{g} - t_0, \zeta - \sqrt{n}(\nu^* - t_0)\rangle_W \\ &= \mu_n\gamma_n(\theta) + m_n\delta_n(\theta) + 2T_{1n} - 2T_{2n} + T_{3n} \end{aligned}$$

where by Lemmas 8 and 9 the terms

$$\begin{aligned} T_{1n} &:= \sqrt{n}\langle \bar{g} - t_0, \zeta\rangle_W \\ T_{2n} &:= \sqrt{n}\langle \bar{g} - t_0, t^* - \mathbb{E}[\hat{t}_n]\rangle_W \\ T_{3n} &:= \|\zeta_n - \sqrt{n}(\hat{t}_n - t_{0n})\|_W^2 - \mathbb{E}\left[\|\zeta_n - \sqrt{n}(\hat{t}_n - t_{0n})\|_W^2\right] \end{aligned}$$

converge to zero uniformly in $\theta \in \Theta$ after normalizing with $h_n(\theta)^{-1}$ □

Proof of Theorem 2:

Note first that Conditions 3 (a) and (b) are clearly satisfied. Now, denote $\hat{\gamma}_n := \mu_n^{-1}(n\hat{Q}_n - \alpha_n\delta_n)$. In order to verify (c), note that we have for any $\varepsilon > 0$ and $\eta > 0$ and $\tau \in \mathbb{R}$ and n large enough, by Assumption 5 and Proposition 1

$$\begin{aligned} 1 - \varepsilon &< P\left(\sup_{\theta\in\Theta}\left|\frac{\hat{\gamma}_n}{\hat{\gamma}_n \vee 1}\frac{\hat{\gamma}_n \vee 1}{\gamma_n \vee 1} - \frac{\gamma_n}{\gamma_n \vee 1}\right| < \eta\right) \\ &= P\left(\sup_{\theta\in\Theta}\left|(\hat{\gamma}_n \vee 1)\frac{(\gamma_n + \tau(\gamma_n \vee 1)) \vee 1}{\gamma_n \vee 1} - (\gamma_n \vee 1)\right| < \eta\right) \\ &\leq P\left(\sup_{\theta\in\Theta}|(\hat{\gamma}_n \vee 1) - (\gamma_n \vee 1)| < \eta + \tau\right) \end{aligned}$$

Since we can choose ε, η, and τ arbitrarily close to zero, so that Condition 3 (c) holds for $K = 1$, and part (d) follows from Assumption 3 part (c). Condition 4 is satisfied by Assumption 1 and Proposition 1 with $\delta = 1$. Finally, Condition 5 has already been shown to hold in Lemma 1 □

Proof of Proposition 2:

By Condition 6 (c), we can w.l.o.g. assume that each component of ζ_n has unit variance. By Condition 7 (d), we have

$$\begin{aligned}\limsup_n P(S(\hat{g}_n, W_n) > S(\varphi_n(\hat{g}_n), \psi_n(W_n))) &= \limsup_n P(S(\varphi(\hat{g}_n), \psi(W_n)) > S(\varphi_n(\hat{g}_n), \psi_n(W_n))) \\ &\leq \limsup_n P(\varphi(\hat{g}_n) \neq \varphi_n(\hat{g}_n)) = 0\end{aligned}$$

since by a Hilbert space version of Strassen's Law of the Iterated Logarithm (HLIL, e.g. Theorem 8.5 in Ledoux and Talagrand (1991) or Theorem 3.1 in Kuelbs and Kurtz (1974)) applied to $n^{1/2}\hat{g}_n - h_1$,

$$\begin{aligned}\limsup_n \quad P \quad &(\hat{g}_{mn} > \varrho_n \text{ for some } m \text{ such that } \eta_{1nm} < \infty) \\ &\leq \limsup_n P\left(\frac{\eta_{1n} + \zeta_{mn}}{\sqrt{2\log\log n}} > 1 \text{ for some } m \text{ such that } \eta_{1nm} < \infty\right) \\ &\leq P\left(\limsup_n \frac{\eta_{1n} + \zeta_{mn}}{\sqrt{2\log\log n}} > 1 \text{ for some } m \text{ such that } \eta_{1nm} < \infty\right) = 0\end{aligned}$$

By the same argument, $\limsup_n P(S(\hat{g}_n, W_n) > S(\varphi_n(\hat{g}_n), \psi_n(W_n))) = 0$ □

Proof of Theorem 3

In order to prove the first statement in Proposition 3, I will first show that

$$\limsup_n \sup_{(\theta,P)\in\mathcal{P}_0} |P_F(T_n(\theta) \leq x) - P(T(\eta_n, \theta) \leq x)| = 0$$

for all $x \geq 0$. Following the argument in Andrews and Guggenberger (2007a), since the limsup over the infimum over (θ, P) has to be attained along some subsequence of (θ_n, η_n), it will be sufficient to verify convergence along any subsequence w_n, or equivalently,

$$\lim_n |P_{\eta_n}(T_n(\theta_n) \leq x) - P(T_n^*(\eta_{1n}, \eta_{2n}, \theta_n) \leq x)| = 0 \tag{A.2}$$

where P_{η_n} denotes the sequence of probability measures in $\mathcal{P}_0$ indexed by the sequence (θ_n, η_n).

By Condition 6 (c) we can w.l.o.g. assume that each component of ζ_n has unit variance, since under $\frac{m_n}{n} \to 0$ we can always pre-multiply the moment vector with a diagonal matrix containing consistent estimators of the (marginal) standard deviations for each element. Since some of the elements in h_1 may be equal to infinity, $h_1 + \zeta_n$ need not be a proper random vector so that it is not possible to apply the Berry-Esséen bound directly to the sequences $\eta_{1n} + \zeta_{w_n}$. I will therefore use the following truncation argument:
Define $\eta^*_{1nm} := \min\{\eta^*_{1n,m}, 2\varrho_n\}$. Clearly, for every n, the sequence $\{\eta^*_{1w_nm}\}_{m\geq 1}$ is in ℓ^∞, the space of bounded sequences. By the definition of the truncated parameter sequences η^*_{1n}, and the argument from the proof of Proposition 2, we have

$$\begin{aligned}
\limsup_n \quad P \quad & (S(\eta_{1n} + \zeta_n, W_n) \neq S(\eta^*_{1n} + \zeta_n, W_n)) \\
\leq \quad & \limsup_n P(S(\eta_{1n} + \zeta_n, W_n) \neq S(\varphi_n(\eta_{1n} + \zeta_n), W_n)) \\
& + \limsup_n P(S(\varphi_n(\eta_{1n} + \zeta_n), W_n) \neq S(\eta^*_{1n} + \zeta_n, W_n)) \\
= \quad & \limsup_n P(S(\varphi(\eta_{1n} + \zeta_n), W_n) \neq S(\varphi_n(\eta_{1n} + \zeta_n), W_n)) \\
& + \limsup_n P(S(\varphi_n(\eta^*_{1n} + \zeta_n), W_n) \neq S(\varphi(\eta^*_{1n} + \zeta_n), W_n)) \\
= \quad & 0
\end{aligned}$$

for any $x \in \mathbb{R}$. Hence it is sufficient to consider the truncated sequences η^*_{1n}.
Since the dimension m_n of the moment vector increases in n, I will use dimension-dependent Berry-Esséen bounds to justify the approximation of the distribution of ζ_n by a Gaussian vector. Since by Condition 6 (d) $S(\gamma + Z, W)$ is quasi-convex in Z, the lower contour sets $\mathrm{con}_{\leq x} S(\cdot, W) := \{g \in l_2 : S(g, W) \leq x\}$ are convex for every $x \in \mathbb{R}$. In particular, if only the first m components of g are nonzero, the projection

of $\mathrm{con}_{\leq x} S$ onto the first m coordinates is convex.

For the class C^m of convex sets in $\mathbb{R}^m$, it follows from Theorem 1.1 in Bentkus (2003) that for an orthonormal Gaussian sequence Z in l_2,

$$\sup_{A \in C^m} \left| P(\pi_m(\xi_n(\theta)) \in A) - P(\pi_m(\Omega(\theta)^{1/2} Z) \in A) \right| \leq \frac{400 m^{1/4} \mathbb{E} \|\xi(\theta)\|^3}{\sqrt{n}}$$

where $\pi_m(x)$ denotes the projection of x onto its first m components. Hence by Assumption 8,

$$\sup_{t \in \mathbb{R}} \left| P_{\eta_n^*}(\hat{T}_n(\theta_n) \leq x) - P(T_n^*(\eta_{1n}^*, \theta_n) \leq x) \right| \leq \frac{400 C_2^{3/2} m^{7/4}}{n^{1/2}}$$

This bound depends only on dimension and sample size, the second moment of the distribution of $g_{in}(\theta)$ and an absolute constant, and therefore holds uniformly in the parameter space.

For any sequence η_n, we now have from the above argument that for a Gaussian vector Z_n with mean zero and covariance operator $\Omega_{n,\eta_{2n}}$,

$$\begin{aligned} P_{\eta_n}(\hat{T}_n(\theta) \leq x) &= P\left(a_m S(\varphi_{nm_n}(\eta_{1n} + \zeta_n), \psi(W_{n,\eta_{2n}}, m_n, n)) + o_P(1) \leq x\right) \\ &= P\left(a_m S(\varphi_{nm_n}(\eta_{1n}^* + Z_n), \psi(W_{n,\eta_{2n}}, m_n, n)) \leq x + o_P(1)\right) + O\left(\frac{m_n^7}{n^2}\right) \\ &\rightarrow P(T_n^*(\eta_{1n}, \eta_{2n}, \theta_n) \leq x) + O\left(\frac{m_n^7}{n^2}\right) \end{aligned}$$

where the first equality uses Condition 6 (a), the second step uses continuity from Condition 6 (b) and continuity of the distribution of $S(h + Z, W)$ from Condition 7 (a). Hence, from the rate restriction from Condition 9,

$$\lim_n \left| P_{\eta_n}(\hat{T}_n(\theta_n) \leq x) - P(T_n^*(\eta_{1n}, \eta_{2n}, \theta_n) \leq x) \right| = 0$$

for all $x > 0$.

Hence, from the definition of the critical value and continuity of the distribution function,

$$\begin{aligned}
\limsup_n \sup_{(\theta,P)\in\mathcal{P}_0} P\left(\hat{T}_n(\theta) > \hat{c}_F(\theta,1-\alpha)\right) &\leq \limsup_n \sup_{(\theta,P)\in\mathcal{P}_0} P\left(T_n^*(0,\hat{\eta}_{2n},\theta) > \hat{c}_F(\theta,1-\alpha)\right) \\
&+ \limsup_n \sup_{\eta\in H} \left\{P\left(T_n^*(\eta_1,\eta_2,\theta) > \hat{c}_F(\theta,1-\alpha)\right) - P\left(\hat{T}_n^*(0,\hat{\eta}_2,\theta) > \hat{c}_F(\theta,1-\alpha)\right)\right\} \\
&+ \limsup_n \sup_{(\theta,P)\in\mathcal{P}_0} \left|P(\hat{T}_n(\theta) \leq x) - P(T_n^*(\eta_{1n},\eta_{2n},\theta_n) \leq x)\right| \\
\leq \quad & \alpha + 0 + 0
\end{aligned}$$

The last inequality follows from the following arguments: since under the null hypothesis H, $\eta \geq 0$, Condition 6 (a) implies $P(T_n^*(\eta_1,\hat{\eta}_2,\theta_n) > \hat{c}_F(\theta,1-\alpha)) \leq P(T_n^*(0,\hat{\eta}_2,\theta_n) > \hat{c}_F(\theta,1-\alpha))$ for every $\eta \in H$. Conditions 8 and 9 together imply that $\hat{\eta}_{2n}$ is consistent for $\eta_2 n$ in the sense that $\sup_{1\leq k,l\leq m_n} |\hat{\eta}_{2n,kl} - \eta_{2n,kl}| \overset{n\to\infty}{\longrightarrow} 0$, where $\eta_{2n,kl}$ denotes the (k,l)th element of η_{2n}. Hence by continuity of S in η_2, the second term on the right-hand side is less than or equal to zero. The third term vanishes by (A.2), and the first term is equal to α by construction.

To prove the second part of the theorem, notice that if Condition 10 holds, there is a probability distribution in $\mathcal{M}$ which attains $h_1 = 0$ at some value θ_0, so that $P_h(\hat{T}_n(\theta_0) \leq \hat{c}_F(\theta_0, 1-\alpha)) = 1-\alpha$, so that by continuity of the c.d.f. at its $1-\alpha$ quantile, $AsyCS = 1-\alpha$ □

Proof of Theorem 4

The proof for the first statement of the Theorem follows an argument similar to that in Theorem 3. The crucial additional step needed to establish uniformly validity of critical values constructed via subsampling consists in showing that

$$\limsup_n \sup_{(\theta,P)\in\mathcal{P}_0} \left|L_{nm_nb_n}(\theta,x) - P\left(\hat{T}_n^*(\gamma_{1n},\gamma_{2n},\theta) \leq x\right)\right| = 0 \qquad \text{(A.3)}$$

for all $x > 0$, where $\gamma_{1n} = b_n^r \bar{g}_n(\theta_n)$.

From the rate restriction in 11 and the uniform bound on third moments in Assumption 8, we can again apply the dimension-dependent Berry-Esséen bound from

Bentkus (2003), so that

$$\sup_{(\theta,P)\in\mathcal{P}_0}\left|P_h(\hat{T}_{nmbj}(\theta) > x) - P\left(T_n^*(\gamma_{1n},\gamma_{2n},\theta) > x\right)\right| \leq \frac{400C_2^{3/2}m^{7/4}}{b^{1/2}}$$

By a standard argument, under Assumption 11, 8, and i.i.d. sampling,

$$\sup_{\theta\in\Theta}\left|L_{nm_nb_n}(\theta,x) - P_h(\hat{T}_{nmbj}(\theta) \leq x)\right| \to 0$$

establishing (A.3).

Hence, using a similar argument as in the proof of Theorem 3, along the sequence $(\theta_n, \eta_n, \gamma_n)$,

$$\limsup_n |P(T_n^*(\gamma_{1n},\gamma_{2n},\theta) \leq x) - L_{nm_nb_n}(\theta,x)| = 0$$

Now it is easy to verify that for all $\varepsilon > 0$ and $h \in H$,

$$P(T_n^*(\gamma_{1n},\gamma_{2n},\theta) \leq c_h(\theta, 1-\alpha) + \varepsilon) > 1-\alpha$$

where $c_h(\theta, 1-\alpha)$ is the $1-\alpha$ quantile of the distribution of $\hat{T}_n(\theta)$ under P_h. Hence along the sequence $(\theta_n, \eta_n, \gamma_n)$, the subsampling critical value satisfies

$$\limsup_n P(T_n^*(\gamma_{1n},\gamma_{2n},\theta) > \hat{c}_S(\theta_n, 1-\alpha)) = \limsup_n P_{\gamma_n}(\hat{T}_n(\theta_n) > \hat{c}_S(\theta_n, 1-\alpha)) = \alpha$$

Since for all F and $\theta \in \Theta_{I,n}$, $\bar{g}_n(\theta) \geq 0$, by Assumption 11, we have $\eta_{1n} \geq \gamma_{1n} \geq 0$ for all $(\theta_n, F_n) \in \mathcal{M}$, so that by Assumption 6(a), $P_{\eta_{1n}}(\hat{T}_n(\theta) > x) \leq P_{\gamma_{1n}}(\hat{T}_n(\theta) > x)$, so that at the subsampling critical value,

$$\limsup_n P_{\gamma_n}\left(\hat{T}_n(\theta_n) > \hat{c}_S(\theta_n, 1-\alpha)\right) \leq \alpha$$

establishing the first conclusion of the theorem.

Finally note that if the first part of Theorem 4 applies and if in addition Condition 12 holds, we can show the second part of the Theorem following analogous steps as in Theorem 1 (b) of Andrews and Guggenberger (2007b) with the only difference that instead of considering the limit of the sequences (which need not exist in the infinite-dimensional case), it is possible to apply the argument at any fixed sample size n □

Proof of Proposition 3:

Before we prove Proposition 3, we will show that Condition 13 implies that the sequence $Z_1, Z_2, \dots$ is strong α-mixing. Recall that the mixing coefficients of a random sequence are defined as

$$\alpha_h \equiv \sup_m \sup_{A_1 \in \mathcal{A}_0^m, A_2 \in \mathcal{A}_{m+h}^\infty} |P(A_1 \cap A_2) - P(A_1)P(A_2)|$$

where A_u^v is the σ-algebra generated by the sequence $Z_{u+1}, \dots, Z_v$.

Lemma 10 *Condition 13 implies that $\{Z_m\}_{m \geq 1}$ is a strong mixing sequence, where α_h has size -1.*

Proof: From the definition of the mixing coefficients α_h, it follows directly that $\alpha_h \leq \varrho_h := \sup_{Y_m, Y_{m+h}} \mathbb{E}[Y_m Y_{m+h}]$ where Y_m and Y_{m+h} are mean zero random variables with unit variance which are measurable with respect to $\mathcal{A}_0^m$ and $\mathcal{A}_{m+h}^\infty$, respectively. From Theorem 1 in Kolmogorov and Rozanov (1960) it follows that for Gaussian sequences, ϱ_h coincides with the supremum of correlation coefficients of linear combinations of $Z_1, \dots, Z_m$ and $Z_{m+h}, \dots$, respectively. For a given choice of coordinate pairs (a_1, b_1) and (a_2, b_2), denote the submatrix $\Omega_{a_1,b_1}^{a_2,b_2} := \{\omega_{kl}\}_{a_1 < k \leq a_2}^{b_1 \leq l < b_2}$. Then,

$$\alpha_h^2 \leq \varrho_h^2 = \sup_{x,y \in l_2} \frac{\langle x, \Omega_{0,m}^{m+h,\infty} y \rangle^2}{\|x\|_{\Omega_{0,m}^{0,m}} \|y\|_{\Omega_{m+h,\infty}^{m+h,\infty}}} \leq \sup_{x,y \in l_2} \frac{\langle x, \Omega_{m+h,2m+h}^{m+h,\infty} y \rangle^2}{\|x\|_{\Omega_{0,m}^{0,m}} \|y\|_{\Omega_{m+h,\infty}^{m+h,\infty}}} o\left(h^{-2}\right)$$

by Condition 13 (ii), since we can always reorder the rows of $\Omega_{0,m}^{m+h,\infty}$ in a way such that for all pairs of new indices (k', l) under the translation by $m + h$ and the old indices (k, l), $|k' - l| \geq |k + m + h - l| + h$. Now define $\tilde{x} := (x_1, \dots, x_m, 0, 0, \dots)$, so that for $\|y\|_{\Omega_{m+h,\infty}^{m+h,\infty}} = 1$, by the Cauchy-Schwarz inequality we have

$$\begin{aligned} \langle x, \Omega_{m+h,2m+h}^{m+h,\infty} y\rangle^2 &= \langle \tilde{x}, \Omega_{m+h,\infty}^{m+h,\infty} y\rangle^2 \leq \|\tilde{x}\|^2_{\Omega_{m+h,\infty}^{m+h,\infty}} \|y\|^2_{\Omega_{m+h,\infty}^{m+h,\infty}} \\ &= \|x\|^2_{\Omega_{m+h,2m+h}^{m+h,2m+h}} \cdot 1 \leq B^2 \end{aligned}$$

Hence, $\alpha_h \leq \varrho_h \leq Bo\left(h^{-1}\right)$, so that $Z_1, Z_2, \ldots$ are strong α-mixing with α of size -1 □

PROOF OF PROPOSITION 3: Define $S_m := \sum_{l=1}^m D_l$. Since for each m, Z_m is a mean zero Gaussian, $\mathbb{E}[D_l] = 0$, so that $\mathbb{E}S_m = m/2$.
By Lemma 10 and the definition of D_l, the sequence $D_1, D_2, \ldots$ is strong α mixing with size -1. Note that a Gaussian random variable has moments to any order, so that by Corollary 3.1 from Wooldridge and White (1988)

$$m^{-1/2}\left(S_m - \frac{m}{2}\right) \xrightarrow{d} N(0, \bar{\sigma}_m^2)$$

where $\sigma_m^2 := \mathrm{Var}(m^{-1/2}S_m)$.
Now note that from standard arguments for the chi-bar square distribution (see e.g. Silvapulle and Sen (2005)), $m^{1/2}\hat{T}_{nm}(\theta) \sim \chi^2_{S_M}$. Since $\mathbb{E}\chi_j^2 = j$ and $\mathrm{Var}(\chi_j^2) = 2j$, by the law of iterated expectations,

$$\begin{aligned} \mathrm{Var}(\chi^2_{S_m}) &= \mathbb{E}\left[\mathbb{E}[\chi_j^2|j = S_m]^2 + \mathrm{Var}(\chi_j^2|j = S_m)\right] - \mathbb{E}[\chi^2_{S_m}] \\ &= \mathbb{E}[S_m^2 + 2S_m] - \mathbb{E}[S_m]^2 = \frac{m^2}{4} + \mathrm{Var}(S_m) + m - \frac{m^2}{4} = m(1 + \bar{\sigma}_m^2) \end{aligned}$$

Since $D_1, D_2, \ldots$ are bounded and strong mixing, we have $\sup_m \bar{\sigma}_m^2 < \infty$, so that $\frac{\mathbb{E}\chi^2_{S_m}}{\mathrm{Var}(\chi^2_{S_m})} = \frac{1}{1+\bar{\sigma}_m^2} < \infty$. Hence, the conclusion follows from Corollary 2.2 in Dykstra (1991) □

Verification of Assumption 6 for the Statistics S_1-S_3:

We will first provide results on quadratic forms as auxiliary lemmas, which will then be used to show Assumption 6 for the statistic S_2.

Lemma 11 *Let $\mathcal{C}_1 \subset \mathcal{C}_2 \subset l_2$ be closed convex cones. If $g \in \mathcal{C}_2$, then for any $\gamma \in l_2$,*

$$S(\gamma, W_1) = \|(\mathrm{id} - \mathbf{P}_{\mathcal{C}_1})\gamma\|_W^2 \geq \|(\mathrm{id} - \mathbf{P}_{\mathcal{C}_2})(\gamma + g)\|_W^2 = S(\gamma + g, W_2)$$

where $\mathbf{P}_{\mathcal{C}}x$ denotes the projection of x onto $\mathcal{C}$ under the norm $\|\cdot\|_W$.

PROOF: By definition,

$$t^* = \mathbf{P}_{\mathcal{C}_1}\gamma = \arg\inf_{t \in \mathcal{C}_1} \|\gamma - t\|_W^2$$

Since $\mathcal{C}_1$ is a closed convex subset of a Hilbert space, the infimum is attained and the arginf is unique (see e.g. section 3.12, Theorem 1 in Luenberger (1969)).
Since $g \in \mathcal{C}_2$, $(t^* + g) \in \mathcal{C}_2$. Therefore,

$$\|(\mathrm{id} - \mathbf{P}_{\mathcal{C}_1})\gamma\|_W^2 = \|(\gamma + g) - (g + t^*)\|_W^2 \geq \min_{\nu \in \mathcal{C}_2} \|(\gamma + g) - t\|_W^2 = \|(\mathrm{id} - \mathbf{P}_{\mathcal{C}_2})(\gamma + g)\|_W^2$$

proving the claim □

Lemma 12 *Let $\mathcal{C} \subset l_2$ be a closed convex cone. Then for any positive definite $W \in \Psi$,*

$$S_2(\gamma, W) = \min_{t \in \mathcal{C}} \|\gamma - t\|_W^2 = \|(\mathrm{id} - \mathbf{P})\gamma\|_W$$

is convex in γ.

PROOF: Let $\gamma_1, \gamma_2 \in l_2$ and define

$$t_1^* := \arg\inf_{t \in \mathcal{C}} \|\gamma_1 - t\|_W, \text{ and } t_2^* := \arg\inf_{t \in \mathcal{C}} \|\gamma_2 - t\|_W$$

Then for any $\lambda \in [0,1]$, $\lambda t_1^* + (1-\lambda)t_2^* \in \mathcal{C}$ since $\mathcal{C}$ is a convex cone. Therefore,

$$\begin{aligned}
S(\lambda\gamma_1 + (1-\lambda)\gamma_2, W) &= \inf_{t\in\mathcal{C}} \|\lambda\gamma_1 + (1-\lambda)\gamma_2 - t\|_W^2 \\
&\leq \|\lambda(\gamma_1 - t_1^*) + (1-\lambda)(\gamma_2 - t_2^*)\|_W^2 \\
&\leq \lambda^2\|\gamma_1 - t_1^*\|_W^2 + (1-\lambda)^2\|\gamma_2 - t_2^*\|_W^2 \\
&\quad +2\lambda(1-\lambda)\|\gamma_1 - t_1^*\|_W\|\gamma_2 - t_2^*\|_W \qquad (A.4) \\
&\leq \lambda S(\gamma_1, W) + (1-\lambda)S(\gamma_2, W)
\end{aligned}$$

where the first inequality follows from the fact that $\lambda t_1^* + (1-\lambda)t_2^* \in \mathcal{C}$, and the second from the triangle inequality since $\|\cdot\|_W$ is a norm on l_2 □

Lemma 13 *Let $\hat{g} = \gamma + Z$. Then for any positive definite $W, \Omega \in \Psi$, the critical value $c_{1-\alpha}(\gamma, W, \Omega)$ of $S(\hat{g}, W)$ is a convex function in γ.*

Proof: Since $S(g, W)$ is convex in g, we have for every realization of Z

$$S(\lambda\gamma_1 + (1-\lambda)\gamma_2 + Z, W) \leq \lambda S(\gamma_1 + Z, W) + S(\gamma_2 + Z, W)$$

Therefore,

$$\lambda S(\gamma_1 + Z, W) + (1-\lambda)S(\gamma_1 + Z, W) \succeq_{FOSD} S(\lambda\gamma_1 + (1-\lambda)\gamma_2 + Z, W)$$

implying that

$$\lambda c_{1-\alpha}(\gamma_1, W, \Omega) + (1-\lambda)c_{1-\alpha}(\gamma_2, W, \Omega) \geq c_{1-\alpha}(\lambda\gamma_1 + (1-\lambda)\gamma_2, W, \Omega)$$

proving the claim □

Lemma 14 *Let $\mathcal{C} \subset l_2$ be a closed convex cone. Then for any bounded, positive* *definite* $W \in \Psi$,

$$S_2(\gamma, W) = \min_{t \in \mathcal{C}} \|\gamma - t\|_W^2 = \|(\mathrm{id} - \mathbf{P})\gamma\|_W$$

is continuous in (γ, W).

PROOF: Let $\delta > 0$ and let (γ_1, γ_2) and (W_1, W_2) be such that

$$\max\{\|\gamma_1 - \gamma_2\|, \|\gamma_1 - \gamma_2\|_{W_1}, \|\gamma_1 - \gamma_2\|_{W_2}\} < \delta$$

and $\|W_1 - W_2\| < \delta$. Also for $i = 1, 2$, let $t_i^* = \arg\inf_{t \in \mathcal{C}} \|\gamma_i - t\|_{W_i}$. Then

$$\begin{aligned}
S_2(\gamma_1, W_1) - S_2(\gamma_2, W_2) &= \|\gamma_1 - t_1^*\|_{W_1}^2 - \|\gamma_2 - t_2\|_{W_2}^2 \\
&\leq \|\gamma_1 - t_2^*\|_{W_1}^2 - \|\gamma_2 - t_2^*\|_{W_1}^2 + \|\gamma_2 - t_2^*\|_{W_1}^2 - \|\gamma_2 - t_2^*\|_{W_2}^2 \\
&\leq \|\gamma_1 - \gamma_2\|_{W_1}^2 + |\langle \gamma_2 - t_2^*, (W_1 - W_2)(\gamma_2 - t_2^*)\rangle| \\
&\leq \|\gamma_1 - \gamma_2\|_{W_1}^2 + \|\gamma_2 - t_2^*\|^2 \|W_1 - W_2\| \\
&\leq \delta(1 + \|\gamma_2 - t_2^*\|^2)
\end{aligned}$$

where the first inequality uses optimality of t_1^*, the second inequality follows from the negative triangle inequality, the next line follows from the Cauchy-Schwarz inequality and the definition of the operator norm. Since γ_2, t_2^* are in l_2, the norm in the last expression is finite. By symmetry, we also have $S_2(\gamma_2, W_2) - S_2(\gamma_1, W_1) \leq \delta(1 + \|\gamma_1 - t_1^*\|^2)$, so that $|S_2(\gamma_2, W_2) - S_2(\gamma_1, W_1)| \leq \delta(1 + \max_{i=1,2} \|\gamma_i - t_i^*\|^2)|$. From the same argument, we also get that

$$\big|\|\gamma_1 - t_1^*\|^2 - \|\gamma_2 - t_2^*\|^2\big| = |S_2(\gamma_1, \mathrm{id}) - S_2(\gamma_2, \mathrm{id})| \leq \delta$$

so that for $\varepsilon > 0$ and fixed $\bar{\gamma}$, we can pick $\delta(\varepsilon, \bar{\gamma}) := 1 \wedge \frac{\varepsilon}{2 + \|\bar{\gamma}\|^2}$ so that $|S_2(\bar{\gamma}, W_2) - S_2(\gamma, W_1)| < \varepsilon$ for all γ with $\max\{\|\gamma - \bar{\gamma}\|, \|\gamma - \bar{\gamma}\|_{\bar{W}}\} < \delta$ and all W with $\|W - \bar{W}\| < \delta$, which establishes continuity □

Proof of Lemma 2:

For S_1, the proof of Condition 6 (a)-(d) is immediate, and property (e) follows directly from Lemma 12 using $W = \text{id}$. For S_2, property (a) follows directly from Lemma 11, property (b) was shown in Lemma 14, and (c) is immediate. Property (d) follows from the fact that S_2 is a norm, and part (e) follows from Lemma 12.

For S_3 note that by restricting the vector a to be in the cone defined by the square root of the weighting operator W, we ensure that all linear combinations will be positive, so that the test function is indeed non-increasing in g. Continuity is immediate from the definition of the test statistic and the fact that the respective norms on the vector and operator spaces of interest derive from the scalar product. Next, since for any p.d. diagonal Δ, we can always pick the matrix square root $(\Delta^{-1}W\Delta^{-1})^{1/2} = W^{1/2}\Delta^{-1}$, so that $\langle a, (\Delta^{-1}W\Delta^{-1})^{1/2}\Delta g\rangle = \langle a, W^{1/2}\Delta^{-1}\Delta g\rangle$, so that part (c) is also satisfied. Part (d) is immediate, and for (e), note that for any two vectors g_1, g_2 and $\lambda \in [0, 1]$, we have by the triangle inequality for $\|\cdot\|_-^2$

$$\begin{aligned} S(\lambda g_1 + (1-\lambda)g_2, W) &= \sup_{a\in A} \|\langle a, W^{1/2}(\lambda g_1 + (1-\lambda)g_2)\rangle\|_-^2 \\ &\leq \sup_{a\in A} \left(\lambda^2\|\langle a, W^{1/2}g_1\rangle\|_-^2 + (1-\lambda)^2\|\langle a, W^{1/2}g_2\rangle\|_-^2\right) \\ &\leq \lambda^2 \sup_{a\in A}\|\langle a, W^{1/2}g_1\rangle\|_-^2 + (1-\lambda)^2 \sup_{a\in A}\|\langle a, W^{1/2}g_2\rangle\|_-^2 \\ &\leq \lambda S(g_1, W) + (1-\lambda)S(g_2, W) \end{aligned}$$

Condition 7 can be verified using the same reasoning as in Andrews and Guggenberger (2007b): If $h_m = \infty$ for all m, all three statistics are equal to zero with probability one, so that part (a) is trivially satisfied. On the other hand, if $h_m \neq \infty$ for some m we can, w.l.o.g. assume that $\|h\| < \infty$ since for all components m with $h_m = \infty$, we can set $h_m + Z_m - \nu_m^*$ equal to zero. Then, $S_i(h + Z, \Omega)$ has full support on the positive real numbers, and continuity follows from quasi-convexity of $S_i(\cdot, \cdot)$ and Theorem 11.1 from Davydov, Lifshits, and Smorodina (1998) (note that their Proposition 11.3 extends from convex to quasi-convex functionals), which establishes parts (a) and (b) of Condition ??. For part (c) notice that we have $S_i(g, W) = 0$ if and only if $g \geq 0$ for $i = 1, 2, 3$. Hence for any Gaussian sequence Z with mean zero, we have $P(S_i(Z, W) \leq 0) \leq P(Z_1 \geq 0) = \frac{1}{2}$. Part (d) clearly holds for all three statistics

□

Derivations for Example 10

In our example, $\hat{T}_n(\theta)$ can be represented as a mixture of chi-squared random variables with different degrees of freedom,

$$\bar{\mu}_{S_{m_n}}(\theta) = \mathbb{E}[\hat{T}_n(\theta)] = \sum_{j=1}^{m_n} 2^{-m_n} \binom{m_n}{j} \mathbb{E}\chi_j^2 = \frac{m_n}{2}$$

Since for the chi-squared distribution with j degrees of freedom $\mathbb{E}\left[(\chi_j^2)^2\right] = \left(\mathbb{E}[\chi_j^2]\right)^2 + \text{Var}(\chi_j^2) = j^2 + 2j$, we can now calculate the second moment by

$$\mathbb{E}[\hat{T}_n(\theta)^2] = \sum_{j=1}^{m_n} 2^{-m_n} \binom{m_n}{j} \mathbb{E}(\chi_j^2)^2 = \sum_{j=1}^{m_n} 2^{-m_n} \binom{m_n}{j} (j^2 + 2j) = \frac{1}{4} m_n(m_n + 1) + m_n$$

We can prove the summation formula $\sum_{j=1}^{m} \binom{m}{j} j^2 = m(m+1)2^{m-2}$ for the last step by induction over m: Clearly the expression is true for $m = 0$, so for $m + 1$, we have by the inductive hypothesis and standard binomial identities

$$\begin{aligned}
\sum_{j=1}^{m+1} \binom{m+1}{j} j^2 &= (m+1)^2 + \sum_{j=1}^{m} \left[\binom{m}{j-1} - \binom{m}{j} \right] j^2 \\
&= (m+1)^2 + \sum_{j=1}^{m} \binom{m}{j} j^2 + \sum_{j=0}^{m} \binom{m}{j-1} (j^2 + 2j + 1) \\
&= m(m+1)2^{m-1} + m2^m + 2^m = (m+1)((m+1)+1)2^{(m+1)-2}
\end{aligned}$$

proving the claim.

Hence,

$$\text{Var}(\hat{T}_n(\theta)) = \frac{1}{4} m_n(m_n + 5) - \frac{1}{4} m_n^2 = \frac{5}{4} m_n$$

so that the ratio $\frac{\bar{\mu}_{m_n}(\theta)}{\bar{\sigma}_J^2(\theta)} = \frac{2}{5} < \infty$. Dykstra (1991) shows that for any mixture of chi-squared distributions $W_n = \chi^2_{J_n}$ where J_n is a integer-valued random variable satisfying $\frac{\mathbb{E}[W_n]}{\mathrm{Var}(W_n)} \leq B < \infty$, the standardization of W_n converges to a standard normal if and only if the distribution of degrees of freedom J_n converges to a normal. In this example, $S_M \sim B\left(M, \frac{1}{2}\right)$, so that asymptotic normality follows from the deMoivre-Laplace theorem □

Proof of Proposition 4:

Define

$$\mu_n = \mu_n(\theta) = 1 + \frac{1}{m_n}\sum_{m=1}^{m_n}\left[n\bar{g}_{mn}(\theta)^2\Phi\left(-\frac{\sqrt{n}\bar{g}_{mn}(\theta)}{\sigma_{mn}(\theta)}\right) + \int_{\sqrt{n}\frac{\bar{g}_{mn}(\theta)}{\sigma_{mn}(\theta)}}^{-\sqrt{n}\frac{\bar{g}_{mn}(\theta)}{\sigma_{mn}(\theta)}} z^2\varphi(z)dz\right]$$

where $\varphi(\cdot)$ is the p.d.f. of a standard normal, and let

$$\bar{\sigma}_n := \bar{\sigma}_{m_n}(\theta)^2 = \mathrm{Var}\left(\frac{1}{\sqrt{m_n}}\sum_{m=1}^{m_n}\left(\frac{\zeta_{mn}}{\sigma_{mn}}\right)^2 \mathbb{1}\left\{\zeta_{mn} \leq \bar{g}_{mn}\right\}\right)$$

Note that under the least favorable hypothesis, $\bar{g}_{mn}(\theta) = 0$ for all $m = 1, 2, \dots$, we have $\mu_n(\theta) = 1$.

For part (i), we can write

$$\sqrt{m_n}S_1(\varphi_{nm_n}(\hat{g}_n), \hat{\Omega}_n^{-1}) = \frac{n}{\sqrt{m_n}}\sum_{m=1}^{m_n}\left[\frac{\bar{g}_{mn}(\theta) + \zeta_{mn}(\theta)}{\hat{\sigma}_{mn}(\theta)}\right]_-^2$$

Now, note that since for any random variable Z with finite second moments,

$$\mathrm{Var}(Z) = \mathbb{E}[Z^2\mathbb{1}\{Z < 0\}] + \mathbb{E}[Z^2\mathbb{1}\{Z \geq 0\}] - \mathbb{E}[Z]^2$$

we have

$$\frac{\mathbb{E}\min\{Z, 0\}^2}{\mathrm{Var}Z} = \frac{\mathbb{E}[Z^2\mathbb{1}\{Z < 0\}]}{\mathrm{Var}(Z)} = \frac{1}{2} - \frac{\mathbb{E}[Z^2\mathbb{1}\{Z \geq 0\}] - \mathbb{E}[Z^2\mathbb{1}\{Z < 0\}]}{2\mathrm{Var}(Z)} + \frac{\mathbb{E}[Z]^2}{2\mathrm{Var}(Z)}$$

which is equal to $\frac{1}{2}$ if Z is distributed symmetrically about zero. Hence,

$$\frac{\mathbb{E}\min\{\hat{g}_n, 0\}^2}{\sigma_{mn}^2} = \frac{1}{2} + \frac{n\bar{g}_n^2 P(\hat{g}_n < 0)}{2\sigma_{mn}(\theta)^2} + \frac{\mathbb{E}[\zeta_{mn}^2 \mathbb{1}\{\hat{g}_n < 0\}] - \mathbb{E}[\zeta_{mn}^2 \mathbb{1}\{\hat{g}_n \geq 0\}]}{2\sigma_{mn}(\theta)^2}$$

Defining $Z_{mn} = \frac{\zeta_{mn}}{\hat{\sigma}_{mn}}$, we have by integrating by parts

$$\begin{aligned}\mathbb{E}\left[Z_{mn}^2 \mathbb{1}\{Z_{mn} < 0\}\right] &= \int_{-\infty}^{0} z^2 dP(Z_{mn} \leq z) = -2\int_{-\infty}^{0} zP(Z_{mn} \leq z)dz \\ &= 2\int_0^{\infty} z\left[\Phi(z) + n^{-1/2}\frac{1}{6}\mathbb{E}[Z_{mn}^3]\varphi(z) + o(n^{-1/2})\right]dz\end{aligned}$$

by an Edgeworth expansion for the studentized mean (see e.g. Hall (1992) section 2), where $\Phi(z)$ denotes the standard normal c.d.f. and $\varphi(z)$ the standard normal density. By an analogous argument for $\mathbb{E}\left[Z_{mn}^2 \mathbb{1}\{Z_{mn} < 0\}\right]$ and using that the normal distribution is symmetric about zero, we therefore have

$$\mathbb{E}\left[\frac{\zeta_{mn}^2}{\sigma_{mn}^2}\mathbb{1}\{\zeta_{mn} < 0\}\right] - \mathbb{E}\left[\frac{\zeta_{mn}^2}{\sigma_{mn}^2}\mathbb{1}\{\zeta_{mn} \geq 0\}\right] = n^{-1/2}\frac{2\mathbb{E}[\zeta_{mn}^3]}{3\sigma_{mn}^3}\varphi(0) + o(n^{-1/2}) = O(n^{-1/2})$$

where the first equality follows from a standard result for the censored normal mean. For $\bar{g}_{mn} \neq 0$, we get by the same line of reasoning that

$$\mathbb{E}\left[\frac{\zeta_{mn}^2}{\sigma_{mn}^2}\mathbb{1}\{\zeta_{mn} < -\sqrt{n}\bar{g}_{mn}\}\right] - \mathbb{E}\left[\frac{\zeta_{mn}^2}{\sigma_{mn}^2}\mathbb{1}\{\zeta_{mn} \geq -\sqrt{n}\bar{g}_{mn}\}\right] = \int_{\sqrt{n}\bar{g}_{mn}}^{-\sqrt{n}\bar{g}_{mn}} z^2\varphi(z)dz + O(n^{-1/2})$$

Hence,

$$\begin{aligned} m_n\mu_n(\theta) &= \mathbb{E}\left[\sum_{m=1}^{m_n}\frac{\min\{\sqrt{n}\bar{g}_{mn} + \zeta_{mn}, 0\}^2}{\sigma_{mn}^2}\right] \\ &= \frac{m_n}{2} + \frac{1}{2}\sum_{m=1}^{m_n}\left[n\bar{g}_{mn}^2\Phi\left(-\frac{\sqrt{n}\bar{g}_{mn}}{\sigma_{mn}}\right) + \int_{\sqrt{n}\bar{g}_{mn}}^{-\sqrt{n}\bar{g}_{mn}} z^2\varphi(z)dz\right] + O\left(\frac{m_n}{n}\right)^{1/2}\end{aligned}$$

where $O\left(\frac{m_n}{n}\right) = o(1)$ by assumption. By the same reasoning as in the proof of Proposition 3, we get the expression of the asymptotic variance up to a term of order

$\sqrt{\frac{m_n}{n}}$. We can now use Corollary 3.1 from Wooldridge and White (1988) to show that

$$\sqrt{m_n}\frac{\frac{n}{m_n}\sum_{m=1}^{m_n}\frac{\min\{\sqrt{n}\bar{g}_{mn}+\zeta_{mn},0\}^2}{\sigma_{mn}^2}-\mu_n}{\sqrt{1+\bar{\sigma}_n^2}} \xrightarrow{d} N(0,1)$$

For part (ii), note that if the variance operator is diagonal, the minimization problem in (ii) simplifies to

$$m_n S_2(\varphi_1(\hat{g}_n, m_n, n), \Omega_n^{-1}) = n \min_{\nu\geq 0}(\hat{g}_n(\theta)-\nu)'\Omega_n(\theta)^{-1}(\hat{g}_n(\theta)-\nu) = \sum_{m=1}^{m_n}\frac{\min\{\sqrt{n}\bar{g}_n(\theta)+\zeta_n(\theta),0\}^2}{\sigma_{mn}(\theta)^2}$$

Hence, the proof is identical to that for part (i). In order to see why under the least favorable hypothesis $\bar{\sigma}_{m_n} = 2$, note that by an argument completely analogous to that for the expectation, we can show that for a standard normal random variable Z,

$$\begin{aligned}
\mathrm{Var}\left(\frac{\zeta_{mn}^2}{\sigma_{mn}^2}\mathbb{1}\{\zeta_{mn}<0\}\right) &= \mathrm{Var}\left(Z^2\mathbb{1}\{Z<0\}\right)+O(n^{-1/2})\\
&= \frac{1}{2}\mathrm{Var}(Z^2|Z<0)+\frac{1}{2}\left(\mathbb{E}[Z^2\mathbb{1}\{Z<0\}|Z\geq 0]-\mathbb{E}[Z^2\mathbb{1}\{Z<0\}]\right)^2\\
&\quad +\frac{1}{2}\left(\mathbb{E}[Z^2\mathbb{1}\{Z<0\}|Z<0]-\mathbb{E}[Z^2\mathbb{1}\{Z<0\}]\right)^2+O(n^{-1/2})\\
&= 1+\frac{1}{8}+\frac{1}{8}+O(n^{-1/2})=\frac{5}{4}+O(n^{-1/2})
\end{aligned}$$

using the conditional variance identity. Hence if Ω_n is diagonal,

$$\mathrm{Var}\left(\frac{1}{\sqrt{m_n}}\sum_{m=1}^{m_n}\min\left\{\frac{\zeta_{mn}}{\sigma_{mn}},0\right\}^2\right)=\frac{5}{4}+O\left(\frac{m_n}{n}\right)^{1/2}$$

and the conclusion follows from a martingale CLT □

Appendix B

Proofs for Chapter 2

Lemma 15 *Under Assumptions 8 and 9, $\hat{\Omega}_n(\theta) - \Omega_0(\theta) \xrightarrow{p} 0$, and $\hat{C}_{nk}(\theta) - C_{0k}(\theta) \xrightarrow{p} 0$, both uniformly in $\theta \in \Theta_I$ for any $k = 1, \ldots, K$.*

Proof of Lemma 15:

By assumption 9, either $\bar{g}_n(\theta) \to g > 0$ or $n^{1/2}\bar{g}_n(\theta) \to h < \infty$. Since $\hat{g}_{mn}(\theta)$ is a sample average of n independent (or weakly dependent) observations,

$$P\left(\limsup_n \left(\varrho_n \frac{\hat{g}_{mn}(\theta)}{\hat{\sigma}_m(\theta)}\right) < 1\right) = \begin{cases} 0 & \text{if } \bar{g}_n(\theta) \to g > 0 \\ 1 & \text{if } n^{1/2}\bar{g}_n(\theta) \to h < \infty \end{cases}, \tag{B.1}$$

by the law of the iterated logarithm, and similarly,

$$P\left(\liminf_n \left(\varrho_n \frac{\hat{g}_{mn}(\theta)}{\hat{\sigma}_m(\theta)}\right) > 1\right) = \begin{cases} 1 & \text{if } \bar{g}_n(\theta) \to g > 0 \\ 0 & \text{if } n^{1/2}\bar{g}_n(\theta) \to h < \infty \end{cases} \tag{B.2}$$

The estimator of the covariance matrix can be rewritten as

$$\begin{aligned} \hat{\Omega}_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}(g(Y_i,\theta) - \psi_{1n}(\theta))(g(Y_i,\theta) - \psi_{1n}(\theta))' \\ &= \frac{1}{n}\sum_{i=1}^{n}(g(Y_i,\theta) - \hat{g}_n(\theta))(g(Y_i,\theta) - \hat{g}_n(\theta))' + (\hat{g}_n(\theta) - \psi_{1n}(\theta))(\hat{g}(\theta) - \psi_{1n}(\theta))' \end{aligned}$$

For the components with $\bar{g}_{mn}(\theta) \to g_m > 0$, by B.2, $P(\hat{g}_{mn}(\theta) \neq \psi_{1mn}(\theta)) \to 0$, so that the contribution to last term vanishes. For the components with $n^{1/2}\bar{g}_{mn}(\theta) \to h < \infty$, we have that, by B.1, $P(\psi_{1mn}(\theta) \neq 0) \to 0$, so that for any $\varepsilon > 0$, by the triangle inequality

$$P(|\hat{g}_{mn}(\theta) - \psi_{1mn}(\theta)| > \varepsilon) \leq P(|\hat{g}_{mn}(\theta) - \bar{g}_{mn}(\theta)| > \varepsilon) + P(|\psi_{1mn}(\theta) - \bar{g}_{mn}(\theta)| > \varepsilon)$$

where the first term goes to zero by a law of large numbers, and the second term vanishes because

$$P(|\psi_{1mn}(\theta) - \bar{g}_{mn}(\theta)| > \varepsilon) \leq P(\psi_{1mn}(\theta) \neq 0) + \mathbb{1}\{|\bar{g}_{mn}(\theta)| > \varepsilon\} \to 0$$

Note that the latter case also includes exactly binding constraints, $\bar{g}_n(\theta) = 0$. Hence, in either case, by assumption 8 (i), $\|\hat{\Omega}_n(\theta) - \Omega_0(\theta)\| \xrightarrow{p} 0$ uniformly in θ.

Similarly, we have for $\hat{C}_{nk}(\theta)$

$$\begin{aligned}
\hat{C}_{nk}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}(g(Y_i, \theta) - \psi_{1n}(\theta))G_k(Y_i, \theta)' \\
&= \frac{1}{n}\sum_{i=1}^{n}(g(Y_i, \theta) - \hat{g}_n(\theta))G_k(Y_i, \theta)' + (\hat{g}_n(\theta) - \psi_{1n}(\theta))G_k(Y_i, \theta)'
\end{aligned}$$

so that by a similar argument as in the previous step, the second term vanishes, and the first term converges in probability to $C_k(\theta)$ uniformly in θ by Assumption 8 □

PROOF OF PROPOSITION 5:

Omitting the argument θ for notational convenience, notice that by definition of $\hat{\Omega}_n$ and $\hat{C}_{nk}$, $\tilde{G}_{kn}(\theta) := G_k(Y_i, \theta) - (g(Y_i, \theta) - \psi_{1n}(\theta))\hat{\Omega}_n^{-1}\hat{C}_{nk}$ is the average residual from an OLS regression of $G(Y_i, \theta)$ on $g(Y_i, \theta) - \psi_{1n}(\theta)$ which is by construction orthogonal to $g(Y_i, \theta) - \psi_{1n}(\theta)$. Since $Y_1, \dots, Y_n$ are i.i.d., we have for every $k = 1, \dots, K$

$$\begin{aligned}
\mathbb{E}[(\hat{g}_n(\theta) - \psi_{1n}(\theta))\hat{D}_{nk}(\theta)'] &= \frac{1}{n}\mathbb{E}[(g(Y_i, \theta) - \psi_{1n}(\theta))(G_k(Y_i, \theta) - (g(Y_i, \theta) - \psi_{1n}(\theta))\hat{\Omega}_n^{-1}\hat{C}_{nk})] \\
&= 0
\end{aligned}$$

Since for every $\theta \in \Theta_I$, $\psi_{1n}(\theta) - \bar{g}(\theta) \to 0$ a.s., $\hat{D}_n(\theta)$ and $\hat{g}_n(\theta)$ are asymptotically uncorrelated at every $\theta \in \Theta_I$.

Since by the same step as in (B.2) and (B.1) either $\psi_{1mn}(\theta) = 0$ or $\psi_{1mn}(\theta) = \hat{g}_{mn}(\theta)$ with probability converging to 1 for all $m = 1, \ldots, M$, by Assumption 7, Lemma 15, and Slutsky's Theorem,

$$\sqrt{n} \begin{pmatrix} \hat{g}_n(\theta) - \bar{g}_n(\theta) \\ \tilde{G}_n(\theta) - \bar{G}_n(\theta) \end{pmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Omega(\theta) & 0 \\ 0 & V(\theta) - C(\theta)'\Omega(\theta)^{-1}C(\theta) \end{bmatrix} \right)$$

for every $\theta \in \Theta_I$ □

Appendix C

Proofs for Chapter 3

First, we will prove a simple lemma which will be used to justify the local approximation for the Wald statistic:

Lemma 1 *Suppose R.1 and R.2 hold, and let θ_n be a sequence such that $\theta_n \to \theta^* \in \partial\Theta_0$. Then $\bar{\theta}_n \equiv \arg\min_{\hat{m}(\theta)=0} ||\theta_n - \theta||^2$ satisfies $\bar{\theta}_n \to \theta^*$.*

Proof of Lemma 1. Since by assumption R.2, the gradient of $\hat{m}(\theta)$ is bounded away from zero, by the implicit function theorem, the set $\{\theta : \hat{m}(\theta) = 0\}$ is locally approximated by a plane, and we can define

$$\tilde{\theta}_n = \arg\min_{\theta:\hat{m}(\theta)=0} ||\theta^* - \theta||^2 = \theta^* + \nabla_\theta m(\theta^*)(\nabla_\theta m(\theta^*)\nabla_\theta m(\theta^*))^{-1}(\hat{m}(\theta^*) - m(\theta^*)) + o_p(1)$$

By R.1 and R.2, $\tilde{\theta}_n - \theta^* = o_p(1)$ so that by the triangle inequality

$$\begin{aligned} ||\bar{\theta}_n - \theta^*|| &\leq ||\bar{\theta}_n - \theta_n|| + ||\theta_n - \theta^*|| \\ &\leq ||\theta_n - \theta^*|| + ||\theta_n - \tilde{\theta}_n|| \\ &\leq 2||\theta_n - \theta^*|| + ||\theta^* - \tilde{\theta}_n|| = o_p(1) \end{aligned}$$

since $\theta_n \to \theta^*$ and $\tilde{\theta}_n - \theta^* = o_p(1)$.

Proof of Theorem 1

PART 1. (*Limit law of $\mathcal{L}_n$.*) Let $G_n = \sqrt{n}(\widehat{m} - m)$. Then

$$\begin{aligned}\mathcal{L}_n = \sup_{\theta \in \Theta_0} \left[\sqrt{n}\widehat{m}(\theta)/s(\theta)\right]_+^2 &= \sup_{\theta \in \Theta_0} \left[(G_n(\theta) + \sqrt{n}m(\theta))/s(\theta)\right]_+^2 \\ &=_d \sup_{\theta \in \Theta_0} \left[(G(\theta) + \sqrt{n}m(\theta))/\sigma(\theta) + o_p(1)\right]_+^2\end{aligned}$$

The steps, apart from the last, immediately follow from Conditions R.1 and R.3. The last step follows from the argument given below. Indeed, take any sequence $\theta_n \in \Theta_0$ such that

$$\sup_{\theta \in \Theta_0} \left[(G(\theta) + \sqrt{n}m(\theta))/\sigma(\theta) + o_p(1)\right]_+^2 = \left[(G(\theta_n) + \sqrt{n}m(\theta_n))/\sigma(\theta_n) + o_p(1)\right]_+^2.$$

In order for this to occur we need to have that

$$\sqrt{n}m(\theta_n)/\sigma(\theta_n) = O_p(1),$$

which is only possible in view of condition R.2 if, for some stochastically bounded sequence of positive random variables $C_n = O_p(1)$,

$$\sqrt{n}d(\theta_n, \partial\Theta_0) \leq C_n.$$

Therefore we conclude that

$$\begin{aligned}&\sup_{\theta \in \Theta_0} \left[(G(\theta) + \sqrt{n}m(\theta))/\sigma(\theta) + o_p(1)\right]_+^2 \\ &= \sup_{\theta \in \partial\Theta_0, \theta + \lambda/\sqrt{n} \in \Theta_0, \|\lambda\| \leq C_n} \left[(G(\theta + \lambda/\sqrt{n}) + \sqrt{n}m(\theta + \lambda/\sqrt{n}))/\sigma(\theta + \lambda/\sqrt{n}) + o_p(1)\right]_+^2\end{aligned}$$

Using stochastic equicontinuity of G and continuity of σ, the last quantity is further approximated by

$$\sup_{\theta \in \partial\Theta_0, \theta + \lambda/\sqrt{n} \in \Theta_0, \|\lambda\| \leq C_n} \left[(G(\theta) + \sqrt{n}m(\theta + \lambda/\sqrt{n}))/\sigma(\theta) + o_p(1)\right]_+^2.$$

Because $\sqrt{n}m(\theta + \lambda/\sqrt{n}) \leq 0$ and $m(\theta) = 0$ for $\theta \in \Theta_0$ and $\theta + \lambda/\sqrt{n} \in \Theta_0$, we conclude that the last quantity is necessarily equal to $\sup_{\theta \in \partial\Theta_0} [G(\theta)/\sigma(\theta)]_+^2$, yielding the conclusion we needed.

PART 2. (*Limit Law of $\mathcal{W}_n$*). We will begin by justifying the approximation holding with probability going to one

$$\sup_{\theta \in \Theta_0} \sqrt{n} d(\theta, \widehat{\Theta}_0) = \sup_{\Theta_n} \sqrt{n} d(\theta, \widehat{\Theta}_0). \tag{C.1}$$

where

$$\Theta_n = \{\theta \in \Theta_0 : \sqrt{n} d(\theta, \partial\Theta_0) \leq C_n\}$$

where C_n is some stochastically bounded sequence of positive random variables, $C_n = O_p(1)$. Note that right hand side is less than or equal to the left hand side in general, so we only need to show that the right hand side can not be less. Indeed, let θ_n be any sequence such that

$$\sup_{\theta \in \Theta_0} \sqrt{n} d(\theta, \widehat{\Theta}_0) = \sqrt{n} d(\theta_n, \widehat{\Theta}_0).$$

If $\widehat{m}(\theta_n) \leq 0$, then $d(\theta_n, \widehat{\Theta}_0) = 0$, and the claim follows trivially since the right hand side of (C.1) is non-negative and is less than or equal to the left hand side of (C.1). If $\widehat{m}(\theta_n) > 0$, then $d(\theta_n, \widehat{\Theta}_0) > 0$, but for this and for $\theta_n \in \Theta_0$ to take place we must have that $0 < \widehat{m}(\theta_n) = O_p(1/\sqrt{n})$, which by Condition R.2 implies that $d(\theta_n, \widehat{\Theta}_0) = O_p(1/\sqrt{n})$.

In the discussion the quantity $\theta^*(\theta)$ as follows

$$\theta^*(\theta) \in \arg\min_{\theta' \in \partial\Theta_0} ||\theta - \theta'||^2.$$

The argmin set $\theta^*(\theta)$ is a singleton simultaneously for all $\theta \in \Theta_n$, provided n is sufficiently large. This follows from condition R.2 imposed on the gradient $\nabla_\theta m$. Moreover, by examining the optimality condition we can conclude that we must have that for $\theta \in \Theta_n$

$$(I - \nabla_\theta m(\theta^*)(\nabla_\theta m(\theta^*)'\nabla_\theta m(\theta^*))^{-1}\nabla_\theta m(\theta^*)')(\theta - \theta^*) = 0 \qquad \text{(C.2)}$$

The projection of $\theta \in \Theta$ onto the set $\widehat{\Theta} := \{\theta \in \Theta : \widehat{m}(\theta) \leq 0\}$ is given by

$$\tilde{\theta}(\theta) = \arg \min_{\theta' : \widehat{m}(\theta') \leq 0} \|\theta - \theta'\|^2.$$

If $\widehat{m}(\theta) \leq 0$, then $\tilde{\theta}(\theta) = \theta$. If $\widehat{m}(\theta) > 0$, then $\tilde{\theta}(\theta) = \bar{\theta}(\theta)$, where

$$\bar{\theta}(\theta) = \arg \min_{\theta' : \widehat{m}(\theta') = 0} \|\theta - \theta'\|^2.$$

In what follows we will suppress the indexing by θ in order to ease the notation, but it should be understood that we will make all the claims uniformly in $\theta \in \Theta_n$. For each θ, the Lagrangian for this problem is $\|\theta - \theta'\|^2 + 2\widehat{m}(\theta')'\lambda$. Therefore, the quantity $\bar{\theta}(\theta)$ can be take to be an interior solution of the saddle-point problem

$$(\bar{\theta} - \theta) + \nabla_\theta \widehat{m}(\bar{\theta})\lambda = 0$$
$$\widehat{m}(\bar{\theta}) = 0$$

The corner solutions do not contribute to the asymptotic behavior of $\mathcal{W}_n$, and thus can be ignored. A formal justification for this will be presented in future versions of this work. By lemma 1, we can use a mean-value expansion to obtain

$$(\bar{\theta} - \theta) + \nabla_\theta \widehat{m}(\bar{\theta})\lambda = 0$$
$$\widehat{m}(\theta^*) - m(\theta^*) + \nabla_\theta m(\check{\theta})'(\bar{\theta} - \theta^*) = 0$$

Using the partitioned inverse formula, we can verify that under the regularity condition R.2, $\lambda = O_P(\theta - \theta^*)$. Also, $\nabla_\theta \widehat{m}(\bar{\theta}) = \nabla_\theta m(\theta^*) + o_p(1)$ and $\nabla_\theta m(\bar{\theta}) =$

$\nabla_\theta m(\theta^*) + o_p(1)$ uniformly in $\theta \in \Theta$, solving for $(\bar{\theta} - \theta)$ we obtain

$$\begin{aligned} \bar{\theta} - \theta^* &= \nabla_\theta m(\theta^*)(\nabla_\theta m(\theta^*)'\nabla_\theta m(\theta^*))^{-1}(\widehat{m}(\theta^*) - m(\theta^*)) \\ &+ (I - \nabla_\theta m(\theta)(\nabla_\theta m(\theta)'\nabla_\theta m(\theta))^{-1}\nabla_\theta m(\theta)')(\theta - \theta^* + o_p(\theta - \theta^*)) \end{aligned}$$

Using that $m(\theta^*) = 0$ and $\sqrt{n}\widehat{m}(\theta^*) =_d G(\theta^*) + o_p(1)$, we obtain

$$\begin{aligned} \sqrt{n}(\bar{\theta} - \theta^*) &=_d \nabla_\theta m(\theta^*)(\nabla_\theta m(\theta^*)'\nabla_\theta m(\theta^*))^{-1}G(\theta^*) \\ &+ \sqrt{n}(I - \nabla_\theta m(\theta^*)(\nabla_\theta m(\theta^*)'\nabla_\theta m(\theta^*))^{-1}\nabla_\theta m(\theta^*)')(\theta - \theta^*) + o_p\left(\sqrt{n}(\theta - \theta^*)\right) \end{aligned}$$

Furthermore, by $\theta \in \Theta_n$ and by the approximate orthgonality condition (C.2) we further have that $(I - \nabla_\theta m(\theta^*)(\nabla_\theta m(\theta^*)'\nabla_\theta m(\theta^*))^{-1}\nabla_\theta m(\theta^*)')(\theta - \theta^*) = 0$, so that

$$\sqrt{n}(\bar{\theta} - \theta^*) =_d \nabla_\theta m(\theta)(\nabla_\theta m(\theta)'\nabla_\theta m(\theta))^{-1}G(\theta) + o_p(1).$$

We next approximate $1(\widehat{m}(\theta) > 0)$ using that

$$\begin{aligned} \sqrt{n}\widehat{m}(\theta) &= \sqrt{n}\widehat{m}(\bar{\theta}) + \nabla_\theta m(\check{\theta})\sqrt{n}(\theta - \bar{\theta}) \\ &= \nabla m(\theta)'\sqrt{n}(\theta - \bar{\theta}) + o_p(1), \\ &= G(\theta) + o_p(1) \end{aligned}$$

for an intermediate value $\check{\theta}$, where we used that $\widehat{m}(\bar{\theta}) = 0$.

Thus, uniformly in $\theta \in \Theta_n$ we have that

$$\begin{aligned} \sqrt{n}d(\theta, \widehat{\Theta}_n) &= \|\bar{\theta} - \theta\|^2 1\{\nabla m(\theta)\sqrt{n}(\theta - \bar{\theta}) > 0 + o_p(1)\} \\ &= |\nabla_\theta m(\theta)'\nabla_\theta m(\theta))^{-1/2}G(\theta)|1\{G(\theta) > 0 + o_p(1)\} \\ &= [\|\nabla_\theta m(\theta)\|^{-1}G(\theta) + o_p(1)]_+ \end{aligned}$$

Therefore, given the initial approximation (C.1) we obtain that

$$\mathcal{W}_n =_d \sup_{\theta \in \partial\Theta_0} [\|\nabla_\theta m(\theta)\|^{-1} G(\theta)]_+ + o_p(1). \tag{C.3}$$

PART 3. (*Continuity of the Limit Distributions*). The continuity of the distribution function $\mathcal{L}$ on $(0, \infty)$ follows from Theorem 11.1 in Davydov, Lifshits, and Smorodina (1998), and the assumption that the covariance function of G is non-degenerate. Probability that $\mathcal{L}$ is greater than zero is equal to the probability that $\max_j \sup_{\theta \in \Theta} G_j(\theta) > 0$ which is greater than the probability that $G_{j'}(\theta') > 0$ for some fixed j' and θ', but the latter is equal to 1/2. Therefore the claim follows. The claim of continuity of the distribution function of $\mathcal{W}$ on $(0, \infty)$ follows similarly. □

Proof of Corollary 1

This corollary immediately follows from the assumed conditions and from the comments given in the main text preceding the statement of Corollary 1. □

Proof of Theorem 2

We have that $Pr_P[\Theta_0 \subseteq R_{LR}] = Pr_P[\mathcal{L}_n \leq \hat{k}(1-\alpha)]$ by the construction of the confidence region. We then have that for any $\alpha < 1/2$ that $k(1-\alpha)$ is a continuity point of the distribution function of $\mathcal{L}$, so that for any sufficiently small ϵ

$$Pr_P[\mathcal{L}_n \leq \hat{k}(1-\alpha)] \leq Pr_P[\mathcal{L}_n \leq k(1-\alpha) + \epsilon] \rightarrow Pr_P[\mathcal{L} \leq k(1-\alpha) + \epsilon]$$
$$Pr_P[\mathcal{L}_n \leq \hat{k}(1-\alpha)] \geq Pr_P[\mathcal{L}_n \leq k(1-\alpha) - \epsilon] \rightarrow Pr_P[\mathcal{L} \leq k(1-\alpha) - \epsilon]$$

Since we can set ϵ as small as we like and $k(1-\alpha)$ is a continuity point of the distribution function of $\mathcal{L}$, we have that

$$Pr_P[\mathcal{L}_n \leq \hat{k}(1-\alpha)] \rightarrow Pr_P[\mathcal{L} \leq k(1-\alpha)] = (1-\alpha).$$

We can conclude similarly for the W-statistic $\mathcal{W}_n$. □

.

Proof of Corollary 2

This corollary immediately follows from the assumed conditions and Corollary 1. □

Proof of Theorem 3

We have that

$$E_{P^*}[\varphi(V^*)] - E_P[\varphi(V)] = o_p(1) \text{ uniformly in } \varphi \in Lip(C(\Theta)).$$

This implies that

$$E_{P^*}[\varphi([V^*]_+)] - E_P[\varphi([V]_+)] = o_p(1) \text{ uniformly in } \varphi \in Lip(C(\Theta)),$$

since the composition $\varphi \circ [\cdot]_+ \in Lip(C(\Theta))$ for $\varphi \in Lip(C(\Theta))$. This further implies that

$$E_{P^*}[\varphi'(\sup_{R_n}[V^*]_+)] - E_P[\varphi'(\sup_{R_n}[V]_+)] = o_p(1) \text{ uniformly in } \varphi' \in Lip(\mathbb{R}),$$

since the composition $\varphi'(\sup_{R_n}[\cdot]_+) \in Lip(C(\Theta))$ for $\varphi' \in Lip(\mathbb{R})$ and R_n denoting any sequence of closed non-empty subsets in Θ. We have that $\widehat{\partial\Theta_0}$ converges to $\partial\Theta_0$ in the Hausdorff distance, so that

$$\begin{aligned} &|E_P[\varphi'(\sup_{\widehat{\partial\Theta_0}}[V]_+) - \varphi'(\sup_{\partial\Theta_0}[V]_+)]| \\ &\leq E[|\sup_{\widehat{\partial\Theta_0}}[V]_+ - \sup_{\partial\Theta_0}[V]_+| \wedge 1] = o_p(1) \text{ uniformly in } \varphi' \in Lip(\mathbb{R}), \end{aligned}$$

since $\sup_{\widehat{\partial\Theta_0}}[V]_+ - \sup_{\partial\Theta_0}[V] = o_p(1)$ by stochastic equicontinuity of the process V. Since metric ρ_K is a proper metric that satisfies the triangle inequality, we have shown that

$$\rho_K(\mathcal{Q}_{\mathcal{S}^*}, \mathcal{Q}_{\mathcal{S}}) = o_p(1).$$

Next, we note that the convergence $\rho_K(\mathcal{Q}_{\mathcal{S}_n}, \mathcal{Q}_{\mathcal{S}}) = o(1)$, for any sequence of laws $\mathcal{Q}_{\mathcal{S}_n}$ of a sequence of random variables S_n defined on probability space $(\Omega', \mathcal{F}', P_n)$ implies the convergence of the distribution function

$$Pr_{\mathcal{Q}_{\mathcal{S}_n}}[\mathcal{S}_n \leq s] = Pr_{\mathcal{Q}_{\mathcal{S}}}[\mathcal{S} \leq s] + o(1)$$

at each continuity point $(0, \infty)$ of the mapping $s \mapsto Pr[\mathcal{S} \leq s]$ and also convergence of quantile functions

$$\inf\{s : Pr_{\mathcal{Q}_{\mathcal{S}_n}}[\mathcal{S}_n \leq s] \geq p\} = \inf\{s : Pr_{\mathcal{Q}_{\mathcal{S}}}[\mathcal{S} \leq s] \geq p\} + o(1)$$

at each continuity point p of the mapping $s \mapsto \inf\{s : Pr_{\mathcal{Q}_{\mathcal{S}}}[\mathcal{S} \leq s] \geq p\}$. Recall from Theorem 1 that the set of continuity points necessarily includes the region $(0, 1/2)$.

By the Extended Continuous Mapping Theorem we conclude that since $\rho_K(\mathcal{Q}_{\mathcal{S}^*}, \mathcal{Q}_{\mathcal{S}}) = o_p(1)$, for any sequence of laws $\mathcal{Q}_{\mathcal{S}^*}$ of random variable $\mathcal{S}^*$ defined on probability space $(\Omega', \mathcal{F}', P^*)$, we obtain the convergence in probability of the distribution function

$$Pr_{\mathcal{Q}_{\mathcal{S}^*}}[\mathcal{S}^* \leq s] = Pr_{\mathcal{Q}_{\mathcal{S}}}[\mathcal{S} \leq s] + o_p(1)$$

at each continuity point $(0, \infty)$ of the mapping $s \mapsto Pr[\mathcal{S} \leq s]$ and also convergence in probability of the quantile functions

$$\inf\{s : Pr_{\mathcal{Q}_{\mathcal{S}^*}}[\mathcal{S}^* \leq s] \geq p\} = \inf\{s : Pr_{\mathcal{Q}_{\mathcal{S}}}[\mathcal{S} \leq s] \geq p\} + o_p(1)$$

at each continuity point p of the mapping $s \mapsto \inf\{s : Pr_{\mathcal{Q}_{\mathcal{S}}}[\mathcal{S} \leq s] \geq p\}$. □

Proof of Corollary 3

In order to prove this corollary it suffices to show that

$$\rho_K(\mathcal{Q}_{\hat{t}'Z^*}, \mathcal{Q}_{t'Z}; C(\Theta)) = o_p(1).$$

Without loss of generality we can take $\sup \|\hat{t}\| \leq 1$ and $\sup \|t\| \leq 1$. The claim will follow from

$$\rho_K(\mathcal{Q}_{\hat{t}'Z^*}, \mathcal{Q}_{t'Z}; C(\Theta)) \leq \rho_K(\mathcal{Q}_{\hat{t}'Z^*}. \mathcal{Q}_{\hat{t}'Z}; C(\Theta)) + \rho_K(\mathcal{Q}_{\hat{t}'Z}, \mathcal{Q}_{\hat{t}'Z}; C(\Theta)) = o_p(1).$$

That $\rho_K(\mathcal{Q}_{\hat{t}'Z^*}, \mathcal{Q}_{\hat{t}'Z}; C(\Theta)) = o_p(1)$ follows immediately from $\rho_K(\mathcal{Q}_{Z^*}, \mathcal{Q}_Z) = o_p(1)$ and from the mapping $\varphi(\hat{t}'\cdot) \in Lip(\mathbb{R}^k)$ (indeed, $|\varphi(\hat{t}'z) - \varphi(t'z)| \leq \sup |\hat{t}'(z - z')| \wedge 1 \leq [(\sup \|t\| \sup \|z - z'\|) \wedge 1] \leq [\sup \|z - z'\| \wedge 1]$. That $\rho_K(\mathcal{Q}_{\hat{t}'Z}, \mathcal{Q}_{\hat{t}'Z}; C(\Theta)) = o_p(1)$ follows because uniformly in $\varphi \in Lip(C(\Theta)$

$$|E[\varphi(\hat{t}'Z)] - \varphi(t'Z)| \leq E[\sup |(\hat{t} - t)'Z| \wedge 1] \leq E[\sup \|\hat{t} - t\| \|Z\| \wedge 1] = o_p(1).$$

Distribution of the Argmin of $d_W(\theta, \cdot)$

Denote $\Theta_0 := \{\theta \in \Theta : m(\theta, \gamma_0) \leq 0\}$. The projection of θ onto the set $\hat{\Theta} := \{\theta \in \Theta : m(\theta, \hat{\gamma}) \leq 0\}$

$$\bar{\theta} = \arg \min_{\theta : m(\theta', \hat{\gamma}) \leq 0} d_W(\theta. \theta')$$

If $m(\theta, \hat{\gamma}) \leq 0$, then $\bar{\theta} = \theta$. If $m(\theta, \hat{\gamma}) > 0$, then we solve

$$\bar{\theta} = \arg \min_{\theta : m(\theta', \hat{\gamma}) = 0} (\theta - \bar{\theta})' W (\theta - \bar{\theta})$$

The Lagrangian for this problem is

$$\mathcal{L} = (\theta - \bar{\theta})' W (\theta - \bar{\theta}) + 2m(\bar{\theta}, \hat{\gamma})\lambda$$

By the mean value theorem, the first-order conditions for the constrained optimization problem are

$$\begin{aligned} W(\bar{\theta}-\theta)+\nabla_\theta m(\bar{\theta},\hat{\gamma})\lambda &= 0 \\ m(\theta^*,\gamma_0)+\nabla_\theta m(\bar{\theta},\bar{\gamma})(\bar{\theta}-\theta^*)+\nabla_\gamma m(\check{\theta},\check{\gamma})(\hat{\gamma}-\gamma_0) &= 0 \end{aligned}$$

Denoting $\bar{M}_\theta = \nabla_\theta m(\bar{\theta},\hat{\gamma})$, $\check{M}_\theta = \nabla_\theta m(\bar{\theta},\bar{\gamma})(\bar{\theta}-\theta^*)$ and $\check{M}_\gamma = \nabla_\gamma m(\check{\theta},\check{\gamma})$, we can rewrite the first-order conditions in matrix form:

$$\begin{bmatrix} W & \bar{M}_\theta \\ \check{M}_\theta & 0 \end{bmatrix} \begin{bmatrix} \bar{\theta}-\theta^* \\ \lambda \end{bmatrix} = \begin{bmatrix} W(\theta-\theta^*) \\ \check{M}_\gamma(\hat{\gamma}-\gamma_0) \end{bmatrix}$$

Solving for $(\bar{\theta}-\theta)$ via the partitioned inverse formula,

$$\begin{aligned} \bar{\theta}-\theta^* &= W^{-1}\bar{M}_\theta(\bar{M}_\theta' W^{-1}\check{M}_\theta)^{-1}\check{M}_\gamma'(\hat{\gamma}-\gamma_0) \\ &\quad +W^{-1/2}(I-W^{-1/2}\check{M}_\theta(\bar{M}_\theta' W^{-1}\check{M}_\theta)^{-1}\bar{M}_\theta' W^{-1/2})W^{1/2}(\theta-\theta^*) \end{aligned}$$

where the second summand is approximately equal to zero if $\theta-\theta^*$ is small. Note also that $m(\theta,\hat{\gamma}) \geq 0$ if and only if the second component of $\bar{\theta}-\theta$ is positive. Consider a local parameter $\theta_n^* = \theta^* - h/\sqrt{n}$, such that $\theta^* = \arg\min_{m(\theta',\gamma_0)\leq 0} d_W(\theta_n^*,\theta')$. Then, by a central limit theorem and the continuous mapping theorem,

$$\sqrt{n}(\bar{\theta}-\theta_n^*) = \left[\sqrt{n}(\bar{\theta}-\theta^*)+h\right]_{+(2)} \xrightarrow{d} \left[W^{-1}M_\theta(M_\theta' W^{-1}M_\theta)^{-1}M_\gamma'\Omega^{1/2}Z+h\right]_{+(2)}$$

where $Z \sim N(0,I_m)$ and $[x]_{+(2)}$ equals the norm of x if the second component of x is positive and zero otherwise.

Therefore the distribution of the Hausdorff distance can be approximated as

$$\begin{aligned} n\min_{m(\theta,\hat{\gamma})\leq 0} d_W(\theta_n^*,\theta) &= \left[h+\bar{Z}\|\Omega^{1/2}M_\gamma\|(M_\theta' W^{-1}M_\theta)^{-1}M_\theta W^{-1}\right]_{+(2)} \\ &\quad \times W\left[W^{-1}M_\theta(M_\theta' W^{-1}M_\theta)^{-1}\|\Omega^{1/2}M_\gamma\|\bar{Z}+h\right]_{+(2)} \\ &= \left\|\left(\bar{Z}+\frac{\|W^{-1/2}M_\theta\|}{\|\Omega^{1/2}M_\gamma\|}h\right)r(\bar{Z},\theta)\right\|^2 \end{aligned}$$

where

$$r(\bar{Z},\theta) = \frac{\|\Omega^{1/2}M_\gamma\|}{\|W^{-1/2}M_\theta\|}\left\{\left[W^{-1}M_\theta(M_\theta'W^{-1}M_\theta)^{-1}\sigma(\theta^*)\bar{Z}+h\right]_2 \geq 0\right\}$$

In order to see that this approximation is uniform in θ note first that the class of functions

$$\mathcal{F} := \left\{\left\{\left[W^{-1}M_\theta(M_\theta'W^{-1}M_\theta)^{-1}\sigma(\theta^*)\bar{Z}+h\right]_2 \geq 0\right\} : \theta^* \in \Theta\right\}$$

is a VC subgraph class.

On the other hand,
$$\mathcal{G} := \left\{\frac{\|W^{-1/2}M_\theta\|}{\|\Omega^{1/2}M_\gamma\|}\right\}$$
is Lipschitz in the parameter:

$$|g_{\theta_1} - g_{\theta_2}| \leq \sup_{\theta\in\Theta} \|\nabla_\theta g(\theta)\| \|\theta_1 - \theta_2\|$$

where for $\gamma(\theta) := \|M_\theta'W^{-1}M_\theta\|$,

$$\nabla_\theta g(\theta) = \frac{\gamma(\theta)\nabla_\theta\sigma(\theta) - \sigma(\theta)\nabla_\theta\gamma(\theta)}{\gamma(\theta)^2}$$

and we can verify that the gradients of $\sigma(\theta)$ and $\gamma(\theta)$ are indeed uniformly bounded over $\theta \in \Theta$:

$$\begin{aligned}
\frac{\partial}{\partial\theta_k}\sigma(\theta) &= \frac{1}{\sigma(\theta)}M_\gamma'\Omega M_{\gamma\theta_k}\\
\frac{\partial}{\partial\theta_k}\gamma(\theta) &= \frac{1}{\gamma(\theta)}\left(M_\theta'W(\theta)^{-1}M_{\theta\theta_k} - M_\theta'W(\theta)^{-1}W_{\theta_k}(\theta)W(\theta)^{-1}M_\theta\right)
\end{aligned}$$

where the additional subscripts denote derivatives with respect to components of the parameter vector θ_k. Since uniformly in $\theta \in \Theta$ the eigenvalues of W^{-1} and Ω are bounded away from zero, and second derivatives are bounded in absolute value from above, $\sup_{\theta\in\Theta}\|\nabla_\theta g(\theta)\| < \infty$ if Θ is bounded.

Therefore for $\mathrm{diam}(\Theta) < \infty$, $\mathcal{F}$ is bounded Donsker, so that by theorem 2.10.6 from van der Vaart and Wellner (Lipschitz transformations), the product $(\mathcal{G} - h)\mathcal{F}$ is bounded Donsker with constant envelope function $\frac{C_1}{B_0\sqrt{\lambda_W^*\lambda_\Omega^*}} < \infty$, where λ_W^* and λ_Ω^* are the smallest eigenvalues of W and Ω, respectively.

Distribution under Local Alternatives

In order to analyze the power of the two tests against local alternatives, define

$$\bar{m}(\theta, \gamma) = m(\theta, \gamma) + \frac{g(\theta)}{\sqrt{n}}$$

for a non-positive function $g(\theta)$ such that the norm of its first derivative is bounded both from above and away from zero uniformly in θ, i.e. there are constants $0 < C_1 < C_2 < \infty$ such that for all θ $C_1 < \|\nabla_\theta g(\theta)\| < C_2$. Let

$$\Theta_A := \{\theta : \bar{m}(\theta, \gamma_0) \leq 0\}$$

Then by the mean value theorem, we can approximate the LR-type statistic

$$\mathcal{C}_n = \sup_{\theta\in\Theta_0}\left[\frac{m(\theta,\hat{\gamma})}{\sigma(\theta)}\right]_+^2 = \sup_{\theta\in\Theta_0}\left[\frac{M_\gamma(\hat{\gamma}-\gamma_0) - \frac{g(\theta)}{\sqrt{n}}}{\sigma(\theta)} + O((\hat{\gamma}-\gamma_0)^2\right]_+^2$$

Since $\sigma(\theta) = \sqrt{M_\gamma'\Omega M_\gamma}$, $\mathcal{C}_n$ converges weakly to

$$n\mathcal{C}_n \rightsquigarrow \sup_{\theta\in\Theta_0}\left[\frac{M_\gamma'\Omega^{1/2}Z - g(\theta)}{\sigma(\theta)}\right]_+^2 = \sup_{\theta\in\Theta_0}\left[\bar{Z} - \frac{g(\theta)}{\|\Omega^{1/2}M_\gamma\|}\right]_+^2$$

where $\bar{Z} \sim N(0,1)$.

For $\theta^* \in \partial\Theta_0$, define $\bar{\theta} = \arg\min_{\bar{m}(\theta',\gamma_0)=0} d_W(\theta', \theta^*)$. By the mean-value theorem, for some $\check{\theta}$, an element-by-element convex combination of θ^* and $\bar{\theta}$, we can write first-order conditions for this problem as

$$\begin{aligned} W(\bar{\theta}-\theta^*)+\nabla_\theta \bar{m}(\bar{\theta},\gamma_0)\lambda &= 0 \\ \nabla_\theta \bar{m}(\check{\theta},\gamma_0)(\bar{\theta}-\theta^*) &= -\frac{g(\theta^*)}{\sqrt{n}} \end{aligned}$$

since by continuity of $m(\cdot)$, $m(\theta^*,\gamma_0)=0$ for any $\theta^*\in\partial\Theta_0$. Solving for $\bar{\theta}$, we get

$$\sqrt{n}(\bar{\theta}-\theta^*) = W^{-1}[\check{M}_\theta+\check{G}/\sqrt{n}]([\bar{M}_\theta+\bar{G}/\sqrt{n}]'W^{-1}[\check{M}_\theta+\check{G}/\sqrt{n}])^{-1}g(\theta^*)$$

where $\check{G}=\nabla_\theta G(\check{\theta})$ and $\bar{G}=\nabla_\theta(\bar{\theta})$. Since $\|\nabla_\theta g(\theta)\|$ is bounded away from zero uniformly in θ, $\bar{\theta}-\theta^*=O(n^{-1/2})$, and we have that

$$\sqrt{n}(\bar{\theta}-\bar{\theta}) = [\sqrt{n}(\theta^*-\bar{\theta})+\sqrt{n}(\bar{\theta}-\theta^*)]_{+(2)} \xrightarrow{d} \left[W^{-1}M_\theta(M_\theta'W^{-1}M_\theta)^{-1}M_\gamma'(\Omega^{1/2}Z-g(\theta^*))\right]_{+2}$$

pointwise. By uniformity of this approximation, the Wald statistic converges to

$$\begin{aligned} n\mathcal{W}_n &\rightsquigarrow \sup_{\theta\in\partial\Theta_0}\left[(\sigma(\theta)\bar{Z}-g(\theta))(M_\theta'W^{-1}M_\theta)^{-1}M_\theta'W^{-1}\right]_2' W\left[W^{-1}M_\theta(M_\theta'W^{-1}M_\theta)^{-1}(\sigma(\theta)\bar{Z}-g(\theta))\right]_2 \\ &= \sup_{\theta\in\partial\Theta_0}\left[\frac{\sigma(\theta)\bar{Z}-g(\theta)}{\|W^{-1/2}M_\theta\|}\left\{\left[W^{-1}M_\theta(M_\theta'W^{-1}M_\theta)^{-1}(\sigma(\theta)\bar{Z}-g(\theta))\right]_2\geq 0\right\}\right]^2 \\ &= \sup_{\theta\in\partial\Theta_0}\left[\left(\bar{Z}-\frac{g(\theta)}{\|\Omega^{1/2}M_\gamma\|}\right)r(\theta,\bar{Z})\right]^2 \end{aligned}$$

where $r(\theta,\bar{Z})=\frac{\sigma(\theta)}{\|W^{-1/2}M_\theta\|}\left\{\left[W^{-1}M_\theta(M_\theta'W^{-1}M_\theta)^{-1}(\sigma(\theta)\bar{Z}-g(\theta))\right]_2\geq 0\right\}$

Bibliography

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Andrews, D., and P. Guggenberger (2007a): "The Limit of Finite-Sample Size and a Problem with Subsampling," working paper, Yale University and UCLA.

——— (2007b): "Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities," working paper, Yale University and UCLA.

Andrews, D., and P. Jia (2008): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," working paper, MIT and Yale University.

Andrews, D., M. Moreira, and J. Stock (2006): "Optimal Two-Sided Invariant Tests for Instrumental Variables Regression," *Econometrica*, 74(3), 715–752.

Andrews, D., and X. Shi (2008): "Inference for Parameters Defined by Conditional Moment Inequalities," working paper, Yale University.

Andrews, D., and G. Soares (2007): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," working paper, Yale University.

Andrews, D., and J. Stock (2006): "Inference with Weak Instruments," in *Blundell, R., W. Newey, and T. Persson, eds., Advances in Economics and Econometrics Vol.3.*

Bajari, P., L. Benkard, and J. Levin (2007): "Estimating Dynamic Models of Imperfect Competition," *Econometrica*, 75(5), 1331–1370.

Bentkus, V. (2003): "On the Dependence of the Berry- Esséen Bound on Dimension," *Journal of Statistical Planning and Inference*, 113, 385–402.

Beresteanu, A., and F. Molinari (2008): "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76(4), 763–814.

Billingsley, P. (1999): *Convergence of Probability Measures.* Wiley, New York.

Bontemps, C., T. Magnac, and E. Maurin (2007): "Set Identified Linear Models," working paper, Toulouse School of Eocnomics, Paris School of Economics.

Bugni, F. (2008): "Bootstrap Inference in Partially Idendified Models," working paper, Northwestern University.

Canay, I. (2007): "EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity," working paper, University of Wisconsin, Madison.

Carrasco, M., and J.-P. Florens (2000): "Generalization of GMM to a Continuum of Moment Conditions," *Econometric Theory*, 16(6), 797–834.

Chao, J., and N. Swanson (2005): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73(5), 1673–1692.

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. Rosen (2008): "Inference on Intersection Bounds," working paper, MIT and UCL.

Chesher, A. (2005): "Nonparametric Identification under Discrete Variation," *Econometrica*, 73(5), 1525–1550.

Cochrane, J. (2005): *Asset Pricing.* Princeton University Press.

Davydov, Y., M. Lifshits, and N. Smorodina (1998): *Local Properties of Distributions of Stochastic Functionals.* American Mathematical Society, Providence, RI.

De Boor, C., and J. Daniel (1974): "Splines with Nonnegative B-Spline Coefficients," *Mathematics of Computation*, 28, 565–568.

Domínguez, M., and I. Lobato (2004): "Consistent Estimation of Models Defined by Conditional Moment Restrictions," *Econometrica*, 72(5), 1601–1615.

Donald, S., and W. Newey (2000): "A Jackknife Interpretation of the Continuous Updating Estimator," *Economics Letters*, 67, 239–243.

Dykstra, R. (1991): "Asymptotic Normality for Chi-Bar Square Distributions," *The Canadian Journal of Statistics*, 19(3), 297–306.

Geyer, C. (1994): "On the Asymptotics of Constrained M-Estimation," *The Annals of Statistics*, 22(4), 1993–2010.

Hall, P. (1992): *The Bootstrap and Edgeworth Expansion.* Springer, New York.

HAN, C., AND P. PHILLIPS (2006): "GMM with Many Moment Conditions," *Econometrica*, 74(1), 147–192.

HANSEN, L., J. HEATON, AND A. YARON (1996): "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business & Economic Statistics*, 14(3), 262–280.

HANSEN, L., AND R. JAGANNATHAN (1991): "Implications of Security Market Data for Models of Dynamic Economies," *Journal of Political Economy*, 99, 225–262.

IMBENS, G., AND C. MANSKI (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.

KHAN, S., AND E. TAMER (2008): "Inference on Randomly Censored Regression Models Using Conditional Moment Inequalities," forthcoming in Journal of Econometrics.

KIM, K. (2008): "Set Estimation and Inference with Models Characterized by Conditional Moment Inequalities," working paper, University of Minnesota.

KLEIBERGEN, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70(5), 1781–1803.

——— (2005): "Testing Parameters in GMM without Assuming that they are Identified," *Econometrica*, 73(4), 1103–1123.

KNIGHT, K. (1999): "Epi-Convergence in Distribution and Stochastic Equi-Semicontinuity," working paper, University of Toronto.

KOLMOGOROV, A., AND Y. ROZANOV (1960): "On Strong Miximng Conditions for Stationary Gaussian Processes," *Theory of Probability and its Applications*, 5(2), 204–208.

KUDO, A. (1963): "A Multivariate Analogue of the One-Sided Test," *Biometrika*, 50(3), 403–418.

KUELBS, J., AND T. KURTZ (1974): "Berry-Esséen Estimates in Hilbert Spaces and an Application to the Law of the Iterated Logarithm," *Ann. Probability*, 2(3), 387–407.

LEDOUX, M., AND M. TALAGRAND (1991): *Probability in Banach Spaces.* Springer, Heidelberg.

LEHMANN, E., AND J. ROMANO (2005): *Testing Statistical Hypotheses.* Springer, New York.

LINTON, O., K. SONG, AND Y. WHANG (2008): "Bootstrap Tests of Stochastic Dominance with Asymptotic Similarity on the Boundary," unpublished manuscript, LSE, UPenn, and Seoul National University.

LUENBERGER, D. (1969): *Optimization by Vector Space Methods.* Wiley & Sons, New York.

MALINVAUD, E. (1980): *Statistical Methods of Econometrics.* North-Holland, Amsterdam.

MANSKI, C., AND J. PEPPER (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68(4), 997–1010.

MANSKI, C., AND E. TAMER (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

MARKOWITZ, H. (1952): "Portfolio Selection," *Journal of Finance*, 7(1), 77–91.

MIKUSHEVA, A. (2007): "Uniform Inference in Autoregressive Models," *Econometrica*, 75(5), 1411–1452.

MOLCHANOV, I. (1998): "A Limit Theorem for Solutions of Inequalities," *Scandinavian Journal of Statistics*, 25, 235–242.

——— (2005): *Theory of Random Sets.* Springer, London.

MOREIRA, M. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71(4), 1027–1048.

NEWEY, W. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58(4), 809–837.

NEWEY, W., AND R. SMITH (2004): "Higher-Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72(1), 219–255.

NEWEY, W., AND F. WINDMEIJER (2008): "GMM with Many Weak Moment Conditions," working paper, MIT and University of Bristol.

NÜRNBERGER, G. (1989): *Approximation by Spline Functions.* Springer, Heidelberg.

PAKES, A., J. PORTER, K. HO, AND J. ISHII (2006): "Moment Inequalities and their Application," working paper, Harvard University.

POLITIS, D., AND J. ROMANO (1994): "Large Sample Confidence Regions based on Subsamples under Minimal Assumptions," *Annals of Statistics*, 22(4), 2031–2050.

POLITIS, D., J. ROMANO, AND M. WOLF (1999): *Subsampling.* Springer, New York.

ROCKAFELLAR, R., AND R. WETS (1998): *Variational Analysis.* Springer, Heidelberg.

ROMANO, J., AND A. SHAIKH (2006): "Inference for the Identified Set in Partially Identified Econometric Models," working paper, Stanford University and University of Chicago.

ROSEN, A. (2008): "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities," *Journal of Econometrics*, 146, 107–117.

SILVAPULLE, M., AND P. SEN (2005): *Constrained Statistical Inference.* John Wiley and Sons, New York.

STOCK, J., AND J. WRIGHT (2000): "GMM with Weak Identification," *Econometrica*, 68(5), 1055–1096.

STOER, J., AND C. WITZGALL (1970): *Convexity and Optimization in Finite Dimensions I.* Springer, Heidelberg.

STOYE, J. (2009): "More on Confidence Intervals for Partially Identified Parameters," *Econometrica (forthcoming).*

VAN DER VAART, A. (1998): *Asymptotic Statistics.* Cambridge University Press, Cambridge.

VAN DER VAART, A., AND J. WELLNER (1996): *Weak Convergence and Empirical Processes.* Springer, New York.

WOOLDRIDGE, J., AND H. WHITE (1988): "Some Invariance Principles and Central Limit Theorems for Dependent Heterogenous Processes," *Econometric Theory*, 4(2), 210–230.