University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 90

A Practical Two-Step Method for Testing Moment Inequalities

Joseph P. Romano, Azeem M. Shaikh and Michael Wolf

Revised version, April 2014

A Practical Two-Step Method for Testing Moment Inequalities *

Joseph P. Romano †
Departments of Statistics and Economics
Stanford University
romano@stanford.edu

Azeem M. Shaikh ‡
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Michael Wolf
Department of Economics
University of Zurich
michael.wolf@econ.uzh.ch

First version: August 2012§
This version: April 2014

Abstract

This paper considers the problem of testing a finite number of moment inequalities. We propose a two-step approach. In the first step, a confidence region for the moments is constructed. In the second step, this set is used to provide information about which moments are "negative." A Bonferonni-type correction is used to account for the fact that with some probability the moments may not lie in the confidence region. It is shown that the test controls size uniformly over a large class of distributions for the observed data. An important feature of the proposal is that it remains computationally feasible, even when the number of moments is large. The finite-sample properties of the procedure are examined via a simulation study, which demonstrates, among other things, that the proposal remains competitive with existing procedures while being computationally more attractive.

KEY WORDS: Bonferonni inequality, bootstrap, moment inequalities, partial identification, uniform validity.

JEL classification codes: C12, C14.

*We thank a co-editor, four anonymous referees, Ivan A. Canay, and Patrik Guggenberger for helpful comments that have improved the exposition of the paper.

†Research supported by NSF Grant DMS-0707085.

‡Research supported by NSF Grant DMS-1227091 and the Alfred P. Sloan Foundation.

§Under the title "A Simple Two-Step Method for Testing Moment Inequalities with an Application to Inference in Partially Identified Models".

1 Introduction

Let $W_i, i = 1, \ldots, n$, be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$ on $\mathbb{R}^k$ and consider the problem of testing

$$H_0 : P \in \mathbf{P}_0 \text{ versus } H_1 : P \in \mathbf{P}_1 \ , \tag{1}$$

where

$$\mathbf{P}_0 = \{P \in \mathbf{P} : \mathbb{E}_P[W_i] \leq 0\} \tag{2}$$

and $\mathbf{P}_1 = \mathbf{P} \setminus \mathbf{P}_0$. Here, the inequality in (2) is intended to be interpreted component-wise and $\mathbf{P}$ is a "large" class of possible distributions for the observed data. Indeed, we will only impose below a mild (standardized) uniform integrability requirement on $\mathbf{P}$. Our goal is to construct tests $\phi_n = \phi_n(W_1, \ldots, W_n)$ of (1) that are uniformly consistent in level, i.e.,

$$\limsup_{n\to\infty} \sup_{P \in \mathbf{P}_0} \mathbb{E}_P[\phi_n] \leq \alpha \tag{3}$$

for some pre-specified value of $\alpha \in (0, 1)$.

In the interest of constructing tests of (1) that not only satisfy (3), but also have good power properties, it may be desirable to incorporate information about which components of $\mathbb{E}_P[W_i]$ are "negative." Examples of tests that incorporate such information implicitly using subsampling include Romano and Shaikh (2008) and Andrews and Guggenberger (2009), whereas examples of tests that incorporate such information more explicitly include the "*generalized* moment selection" procedures put forward by Andrews and Soares (2010), Canay (2010), and Bugni (2011). Andrews and Barwick (2012a) propose a refinement of "generalized moment selection" termed "*recommended* moment selection" and discuss four reasons why such an approach is preferable. Therefore, our theoretical and numerical comparisons will be mainly restricted to the method of Andrews and Barwick (2012a); extensive comparisons with previous methods are already available in that paper.

Our two-step solution to this problem is similar in spirit to the recommended moment selection approach. In the first step, we construct a confidence region for $\mathbb{E}_P[W_i]$ at some "small" significance level $\beta \in [0, \alpha]$. In the second step, we then use this set to provide information about which components of $\mathbb{E}_P[W_i]$ are "negative" when constructing tests of (1). Importantly, similar to the approach of Andrews and Barwick (2012a), we account in our asymptotic framework for the fact that with some probability, $\mathbb{E}_P[W_i]$ may not lie in the confidence region, using a Bonferonni-type correction; see Remark 2.4 for further discussion.

Our testing procedure and those just cited are related to Hansen (2005), who uses a similar two-stage approach for the same problem, but does not account for the fact that with some probability, $\mathbb{E}_P[W_i]$ may not lie in the confidence region. He instead assumes that β tends to zero as n tends to infinity and only establishes that his test is pointwise consistent in level instead of the stronger requirement (3). The importance of the distinction between (3) and this weaker requirement has been emphasized in the recent literature on inference in partially identified models; for example, see Imbens and Manski (2004), Romano and Shaikh (2008), and Andrews and Guggenberger (2010). Another important feature of our approach stems from our choice of confidence region for $\mathbb{E}_P[W_i]$. Through an appropriate choice of confidence region for $\mathbb{E}_P[W_i]$, our approach remains computationally feasible even when the number of components of $\mathbb{E}_P[W_i]$, denoted by k, is large. In particular, unlike Hansen (2005), we are able to avoid having to optimize over the confidence region numerically.

As described in Remark 2.6, similar computational problems are also present in the approach put forward by Andrews and Barwick (2012a). As a result, they employ computational shortcuts whose validity is only justified using simulation. Even using these shortcuts, they must restrict attention to situations in which $k \leq 10$, which precludes many economic applications, including entry models, as in Ciliberto and Tamer (2009), where $k = 2^{m+1}$ when there are m firms, or dynamic models of imperfect competition, as in Bajari et al. (2007), where k may even be as large as 500. For situations in which $k \leq 10$ and $\alpha = .05$, both procedures are equally easy to implement; however, for situations in which $\alpha \neq .05$, our procedure is considerably easier to implement even when $k \leq 10$. This feature allows us, for example, to construct p-values more easily than Andrews and Barwick (2012a). On the other hand, in contrast to Andrews and Barwick (2012a), we are unable to establish that the lefthand-side of (3) equals α and expect that it is strictly less that α, though we can argue it is not much less than α; see Remark 2.2. Even so, for the situations when both procedures are available, we find in a simulation study that our procedure is nearly as powerful as the one proposed by Andrews and Barwick (2012a).

Other related literature includes Loh (1985), who also uses a similar two-stage approach in the context of some parametric hypothesis testing problems, but, like Hansen (2005), does not account for the fact that with some probability the nuisance parameter may not lie in the confidence region. It is also related to Berger and Boos (1994) and Silvapulle (1996), who improve upon Loh (1985) by introducing a Bonferonni-type correction similar to ours. This idea has been used by Stock and Staiger (1997) to construct a confidence region for the parameters of a linear regression with possibly "weak" instrumental variables. It has also been used in a nonparametric context by Romano and Wolf (2000) to construct a confidence interval for a univariate mean that has finite-sample validity and is "efficient" in a precise sense. Finally, this idea is introduced in a general setting by McCloskey (2012), though the assumptions there technically preclude moment inequality problems; see McCloskey (2012, Section 2.1.3) for further discussion.

The remainder of the paper is organized as follows. In Section S.1 of the online supplement to this paper, we first consider the testing problem in the simplified setting where $\mathbf{P} = \{N(\mu, \Sigma) : \mu \in \mathbb{R}^k\}$ for a known covariance matrix Σ. Here, it is possible to illustrate the main idea behind our construction more clearly and also to obtain some exact results. In particular, we establish an upper bound on the power function of any level-α test of (1) by deriving the most powerful test against any fixed alternative. This result confirms the bound suggested by simulation in Andrews and Barwick (2012b, Section 7.3). We consider the more general, nonparametric setting in Section 2. We apply our main results to the problem of constructing confidence regions in partially identified models defined by a finite number of moment inequalities in Section 3. Section 4 sheds some light on the behavior of our procedures in finite samples via a simulation study, including an extensive comparison of our procedure with the one proposed recently by Andrews and Barwick (2012a). Proofs of all results can be found in the Appendix.

2 The Nonparametric Multi-Sided Testing Problem

Let $W_i, i = 1, \ldots, n$, be an i.i.d. sequence of random vectors with distribution $P \in \mathbf{P}$ on $\mathbb{R}^k$ and consider the problem of testing (1). The unknown family of distributions $\mathbf{P}$ will be a nonparametric class of distributions defined by a mild (standardized) uniform integrability condition, as described in the main results below. Before proceeding, we introduce some useful notation. Below, $\hat{P}_n$ denotes the empirical distribution of the $W_i, i = 1, \ldots, n$. The notation $\mu(P)$ denotes the mean of P and $\mu_j(P)$ denotes the jth component of $\mu(P)$. Let $\bar{W}_n = \mu(\hat{P}_n)$ and $\bar{W}_{j,n} = \mu_j(\hat{P}_n)$. The notation $\Sigma(P)$ denotes the covariance matrix of P and $\sigma_j^2(P)$ denotes the variance of the jth component of P. The notation $\Omega(P)$ denotes the correlation matrix of P. Let $\hat{\Omega}_n = \Omega(\hat{P}_n)$ and $S_{j,n}^2 = \sigma_j^2(\hat{P}_n)$. Finally, let $S_n^2 = \text{diag}(S_{1,n}^2, \ldots, S_{k,n}^2)$.

Our methodology incorporates information about which components of $\mu(P)$ are "negative" by first constructing a (nonparametric) upper confidence rectangle for μ at nominal level $1 - \beta$. Our bootstrap confidence region for this purpose is given by

$$M_n(1-\beta) = \left\{ \mu \in \mathbb{R}^k : \max_{1 \le j \le k} \frac{\sqrt{n}(\mu_j - \bar{W}_{j,n})}{S_{j,n}} \le K_n^{-1}(1-\beta, \hat{P}_n) \right\} , \tag{4}$$

where

$$K_n(x, P) = P\left\{ \max_{1 \le j \le k} \frac{\sqrt{n}(\mu_j(P) - \bar{W}_{j,n})}{S_{j,n}} \le x \right\} . \tag{5}$$

Next, a test statistic T_n is required such that large values of T_n provide evidence against H_0. For simplicity, below we consider several different test statistics of the form

$$T_n = T\left(S_n^{-1}\sqrt{n}\bar{W}_n, \hat{\Omega}_n \right)$$

for some function $T : \mathbb{R}^k \times (\mathbb{R}^k)^2 \to \mathbb{R}$ that is continuous in both arguments and weakly increasing in each component of its first argument. As in Andrews and Barwick (2012a), other test statistics may be considered as well. In particular, we consider

$$T_n^{\max} = \max_{1 \le j \le k} \frac{\sqrt{n}\bar{W}_{j,n}}{S_{j,n}} \tag{6}$$

$$T_n^{\text{qlr}} = \inf_{t \in \mathbb{R}^k : t < 0} Z_n(t)' \hat{\Omega}_n^{-1} Z_n(t) \ , \tag{7}$$

where

$$Z_n(t) = \left(\frac{\sqrt{n}(\bar{W}_{1,n} - t)}{S_{1,n}}, \ldots, \frac{\sqrt{n}(\bar{W}_{k,n} - t)}{S_{k,n}} \right)$$

and the inequality in the infimum is interpreted component-wise. Following Andrews and Barwick (2012a), we also consider an "adjusted" version of T_n^{qlr} in which $\hat{\Omega}_n$ is replaced with

$$\tilde{\Omega}_n = \max\{\epsilon - \det(\hat{\Omega}_n), 0\} \cdot I_k + \hat{\Omega}_n \ ,$$

for some fixed $\epsilon > 0$, with I_k denoting the k-dimensional identity matrix, i.e.,

$$T_n^{\text{qlr,ad}} = \inf_{t \in \mathbb{R}^k : t < 0} Z_n(t)' \tilde{\Omega}_n^{-1} Z_n(t) \ . \tag{8}$$

This modification accommodates situations in which $\Omega(P)$ may be singular. Finally, we also consider the "modified method of moments" test statistic of Andrews and Soares (2010) defined as

$$T_n^{\text{mmm}} = \sum_{j=1}^{k} \left(\frac{\sqrt{n}\bar{W}_{j,n}}{S_{j,n}} \right)^2 \cdot \mathbb{1}\{\bar{W}_{j,n} > 0\} \ . \tag{9}$$

We also require a critical value with which to compare T_n. For $x \in \mathbb{R}$ and $\lambda \in \mathbb{R}^k$, let

$$J_n(x, \lambda, P) = P\left\{ T\left(S_n^{-1}\left(\sqrt{n}(\bar{W}_n - \mu(P)\right) + S_n^{-1}\sqrt{n}\lambda, \hat{\Omega}_n \right) \le x \right\} \ . \tag{10}$$

Note that

$$P\{T_n \le x\} = J_n\big(x, \mu(P), P\big) \ . \tag{11}$$

Importantly, for any x and P, $J_n(x, \lambda, P)$ is nonincreasing in each component of λ. It is natural to replace P in the righthand-side of (11) with $\hat{P}_n$, but this approximation to the distribution of T_n fails when P is on the "boundary" of the null hypothesis; for example, see Andrews (2000). On the other hand, if $\mu(P)$ were known exactly, then one could plug in this value for $\mu(P)$ and replace the final P in the right-hand side of (11) with $\hat{P}_n$. Obviously, $\mu(P)$ is not known exactly, but we may use the confidence region for $\mu(P)$ defined in (4) to limit the possible values for $\mu(P)$. This idea leads us to consider the critical value defined by

$$\hat{c}_n(1 - \alpha + \beta) = \sup_{\lambda \in M_n(1-\beta) \cap \mathbb{R}_-^k} J_n^{-1}(1 - \alpha + \beta, \lambda, \hat{P}_n) \ , \tag{12}$$

where $\mathbb{R}_- = (-\infty, 0]$. The addition of β to the quantile is necessary to account for the possibility that $\mu(P)$ may not lie in $M_n(1 - \beta)$. It may be removed by allowing β to tend to zero with the sample size. However, the spirit of this paper, as well as Andrews and Barwick (2012a), is to account for the selection of moments in order to achieve better finite-sample size performance; see Remark 2.4 below for further discussion.

The calculation of $\hat{c}_n(\cdot)$ in (12) is straightforward because $J_n^{-1}(1-\alpha+\beta,\lambda,\hat{P}_n)$ is nondecreasing in each component of λ. It follows that the supremum in (12) is attained when $\lambda=\lambda^*$ has jth component equal to the minimum of zero and the upper confidence bound for the μ_j, i.e.,

$$\lambda_j^* = \min\left\{\bar{W}_{j,n} + \frac{S_{j,n}K_n^{-1}(1-\beta,\hat{P}_n)}{\sqrt{n}}, 0\right\} \ . \tag{13}$$

Then,

$$\hat{c}_n(1-\alpha+\beta) = J_n^{-1}(1-\alpha+\beta,\lambda^*,\hat{P}_n) \ . \tag{14}$$

Since $\beta\in(0,\alpha)$, we define our test so that it fails to reject the null hypothesis not only whenever T_n is less than or equal to the critical value defined above, but also whenever $M_n(1-\beta)\subseteq\mathbb{R}_-^k$. Formally, our test is, therefore, given by

$$\phi_n = \phi_n(\alpha,\beta) = 1 - \mathbb{1}\left\{\left\{M_n(1-\beta)\subseteq\mathbb{R}_-^k\right\}\cup\left\{T_n\le\hat{c}_n(1-\alpha+\beta)\right\}\right\} \ , \tag{15}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. The following theorem shows that this test controls the probability of a Type I error uniformly over $\mathbf{P}$ in the sense that (3) holds, as long as $\mathbf{P}$ satisfies a mild (standardized) uniform integrability condition.

Theorem 2.1. *Let $W_i, i=1,\dots,n$, be an i.i.d. sequence of random vectors with distribution $P\in\mathbf{P}$ on $\mathbb{R}^k$. Suppose $\mathbf{P}$ is such that, for all $1\le j\le k$,*

$$\lim_{\lambda\to\infty}\sup_{P\in\mathbf{P}}\mathbb{E}_P\left[\left(\frac{W_{j,1}-\mu_j(P)}{\sigma_j(P)}\right)^2\mathbb{1}\left\{\left|\frac{W_{j,1}-\mu_j(P)}{\sigma_j(P)}\right|>\lambda\right\}\right]=0 \ . \tag{16}$$

Fix $0\le\beta\le\alpha$. The test ϕ_n of (1) defined by (15) with T_n given by (6), (8) or (9) satisfies (3).

Remark 2.1. If, in addition to satisfying the requirements of Theorem 2.1, $\mathbf{P}$ is required to satisfy $\inf_{P\in\mathbf{P}}\det(\Omega(P))>0$, then the conclusion of Theorem 2.1 holds when T_n is given by (7). ■

Remark 2.2. By arguing as in Remark S.1.2 in the online supplement to this paper, it is in fact possible to show that the left-hand side of (3) is at least $\alpha-\beta$, so that for small β, the test is not overly conservative. ■

Remark 2.3. In some cases, the null hypothesis may be such that some components of $\mathbb{E}_P[W_i]$ are equal to zero rather than less than or equal to zero. That is, rather than testing that P belongs in $\mathbf{P}_0$ given by (2), the problem is to test that P belongs to $\tilde{\mathbf{P}}_\mathbf{0}$ given by

$$\tilde{\mathbf{P}}_\mathbf{0} = \{P\in\mathbf{P}: \ \mathbb{E}_P[W_{j,1}]=0 \text{ for } j\in J_1, \ \mathbb{E}_P[W_{j,1}]\le 0 \text{ for } j\in J_2\} \ ,$$

where J_1 and J_2 form a partition of $\{1, \ldots, k\}$. Such a situation may be accommodated in the framework described above by writing $\mathbb{E}_P[W_{j,1}] = 0$ as two inequalities $\mathbb{E}_P[W_{j,1}] \le 0$ and $-\mathbb{E}_P[W_{j,1}] \le 0$. Note that it may be possible to improve upon this approach by exploiting the additional structure of the null hypotheses, as is done in Remark S.1.4 in the online supplement to this paper. ■

Remark 2.4. For $\beta = \beta_n$ tending to zero, it follows from our analysis that the test $\phi_n^*(\beta_n)$, where

$$\phi_n^*(\beta) = 1 - \mathbb{1}\Big\{\{M_n(1-\beta) \subseteq \mathbb{R}_-^k\} \cup \{T_n \le \hat{c}_n(1-\alpha)\}\Big\} ,$$

satisfies

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} \mathbb{E}_P\big[\phi_n^*(\beta_n)\big] \le \alpha$$

under the assumptions of Theorem 2.1. To see this, suppose that the assumptions of Theorem 2.1 hold. Let $\phi_n = \phi_n(\alpha, \beta)$ be defined as in (15). Fix any $\epsilon > 0$. By monotonicity, we have for all large enough n that $M_n(1-\beta_n) \subseteq M_n(1-\epsilon)$. Hence, for all such n, we have that $\phi_n^*(\beta_n) \le \phi_n^*(\epsilon)$. Moreover, $\phi_n(\alpha+\epsilon, \epsilon) = \phi_n^*(\epsilon)$. It, therefore, follows from Theorem 2.1 that

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} \mathbb{E}_P\big[\phi_n^*(\beta_n)\big] \le \limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} \mathbb{E}_P\big[\phi_n(\alpha+\epsilon, \epsilon)\big] \le \alpha + \epsilon .$$

Since the choice of $\epsilon > 0$ was arbitrary, the desired result follows. The test $\phi_n^*(\beta_n)$ defined in this way is similar to the "generalized moment selection" procedures of Andrews and Soares (2010), Canay (2010), and Bugni (2011). On the other hand, the test ϕ_n defined by (15), which accounts for the impact of the choice of β on the finite-sample behavior of the testing procedure, is more similar to the procedure of Andrews and Barwick (2012a). ■

Remark 2.5. An "optimal" approach to choosing β is described in Remark S.1.6. We have found that a reasonable simple choice is $\beta = \alpha/10$. Further discussion is given in Section 4.

Remark 2.6. For the hypothesis testing problem considered in this section, Andrews and Barwick (2012a) consider an alternative testing procedure that they term "recommended moment selection." In order to describe a version of their method based on the bootstrap, fix $\kappa < 0$. Let $\hat{\lambda}_n$ be the k-dimensional vector whose jth component equals zero if $\sqrt{n}\bar{W}_{j,n}/S_{j,n} > \kappa$ and $-\infty$ otherwise (or, for practical purposes, some very large negative number). Define the "size correction factor"

$$\hat{\eta}_n = \inf\Big\{\eta > 0 : \sup_{\lambda\in\mathbb{R}^k : \lambda\le 0} J_n(J_n^{-1}(1-\alpha, \hat{\lambda}_n, \hat{P}_n) + \eta, \lambda, \hat{P}_n) \ge \alpha\Big\} . \quad (17)$$

The proposed test is then given by

$$\phi_n(\alpha) = \mathbb{1}\{T_n > J_n^{-1}(1-\alpha, \hat{\lambda}_n, \hat{P}_n) + \hat{\eta}_n\} ,$$

where T_n is given by T_n^{qlr} or $T_n^{\text{qlr,ad}}$; see (7) and (8). The addition of $\hat{\eta}_n$ is required because, in order to allow the asymptotic framework to better reflect the finite-sample situation, the authors do not allow κ to tend to zero with the sample size n. As explained in Remark S.1.5 in the online supplement to this paper, determination of $\hat{\eta}_n$ defined in (17) is computationally prohibitive, even in a parametric setting. This remains true here, so the authors resort to an approximation to the supremum in (17) analogous to the one described in Remark S.1.5. The authors provide an extensive simulation study, but no proof, in favor of this approximation and restrict attention to situations in which $k \leq 10$ and $\alpha = .05$. The authors also provide simulation-based evidence to support a further approximation to $\hat{\eta}_n$ that only depends on k and the smallest off-diagonal element of $\hat{\Omega}_n$. A data-dependent way of choosing κ similar to the way of choosing β described in Remark S.1.6 is described as well. ■

3 Confidence Regions for Partially Identified Models

In this section, we consider the related problem of constructing a confidence region for identifiable parameters that is uniformly consistent in level. Concretely, let $X_i, i = 1, \ldots, n$, be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$ on some general sample space $\mathcal{S}$, where $\mathbf{P}$ is again a nonparametric class of distributions defined by a mild (standardized) uniform integrability requirement on $\mathbf{P}$. We consider the class of partially identified models in which the *identified set*, $\Theta_0(P)$, is given by

$$\Theta_0(P) = \{\theta \in \Theta : \mathbb{E}_P[g(X_i, \theta)] \leq 0\} \ , \tag{18}$$

where Θ is some parameter space (usually some subset of Euclidean space) and $g : \mathcal{S} \times \Theta \to \mathbb{R}^k$. Here, for each θ, $g(\cdot, \theta)$ is a vector of k real-valued functions, and the inequality in (18) is intended to be interpreted component-wise. We wish to construct random sets $\mathcal{C}_n = \mathcal{C}_n(X_1, \ldots, X_n)$ satisfying

$$\liminf_{n \to \infty} \inf_{P \in \mathbf{P}} \inf_{\theta \in \Theta_0(P)} P\{\theta \in \mathcal{C}_n\} \geq 1 - \alpha \tag{19}$$

for some pre-specified $\alpha \in (0, 1)$. As in Romano and Shaikh (2008), we refer to such sets as confidence regions for identifiable parameters that are uniformly consistent in level. Note that in this paper we will not consider the construction of confidence regions for the identified set itself; see Chernozhukov et al. (2007), Bugni (2010), and Romano and Shaikh (2010) for further discussion of such confidence regions.

As in Romano and Shaikh (2008), our construction will be based upon the duality between constructing confidence regions and hypothesis tests. Specifically, we will consider tests of the null hypotheses

$$H_\theta : \mathbb{E}_P[g(X_i, \theta)] \leq 0 \tag{20}$$

for each $\theta \in \Theta$ that control the usual probability of a Type I error at level α. To this end, for each $\theta \in \Theta$, let $\phi_n(\theta)$ be the test of (20) given by the following algorithm.

Algorithm 3.1.

(a) *Set* $W_i = g(X_i, \theta)$.

(b) *Compute the bootstrap quantile* $K_n^{-1}(1-\beta, \hat{P}_n)$, *where* $K_n(x, P)$ *is given by* (5).

(c) *Using* $K_n^{-1}(1-\beta, \hat{P}_n)$ *from (b), compute* $M_n(1-\beta)$ *via equation* (4).

(d) *Using* $K_n^{-1}(1-\beta, \hat{P}_n)$ *from (b), compute* λ^* *via equation* (13).

(e) *Compute the bootstrap quantile* $\hat{c}_n(1-\alpha+\beta) = J_n^{-1}(1-\alpha+\beta, \lambda^*, \hat{P}_n)$, *where* $J_n(x, \lambda, P)$ *is given by* (10).

(f) *Compute* $\phi_n(\theta) = \phi_n$, *where* ϕ_n *is given by* (15).

Consider

$$C_n = \{\theta \in \Theta : \phi_n(\theta) = 0\} \ . \tag{21}$$

The following theorem shows that C_n satisfies (19). In the statement of the theorem, we denote by $\mu_j(\theta, P)$ and $\sigma_j^2(\theta, P)$ the mean and variance, respectively, of $g_j(X_i, \theta)$ under P.

Theorem 3.1. *Let* $X_i, i = 1, \ldots, n$, *be an i.i.d. sequence of random variables with distribution* $P \in \mathbf{P}$. *Suppose* $\mathbf{P}$ *is such that, for all* $1 \leq j \leq k$,

$$\lim_{\lambda \to \infty} \sup_{P \in \mathbf{P}} \sup_{\theta \in \Theta_0(P)} \mathbb{E}_P \left[\left(\frac{g_j(X_i, \theta) - \mu_j(\theta, P)}{\sigma_j(\theta, P)} \right)^2 \mathbb{1} \left\{ \left| \frac{g_j(X_i, \theta) - \mu_j(\theta, P)}{\sigma_j(\theta, P)} \right| > \lambda \right\} \right] = 0 \ .$$

Then, C_n *defined by* (21) *with* T_n *given by* (6), (8) *or* (9) *satisfies* (19).

4 Simulation Study

The goal of this section is to study the finite-sample performance of our two-step procedure. For the reasons mentioned in the introduction, the comparison with other procedures is reserved to the newly-recommended procedure of Andrews and Barwick (2012a) (henceforth abbreviated as AB) with certain details provided in Andrews and Barwick (2012b). In their notation, the preferred procedure is the "recommended moment selection" (RMS) test based on $(S_2, \varphi^{(1)})$ with data-dependent tuning parameters $\widehat{\kappa}$ and $\widehat{\eta}$ and it is termed "qlr, ad/t-Test/κAuto".

We compare finite-sample performance both in terms of maximum null rejection probability (MNRP) and average power for a nominal level of $\alpha = 0.05$. The design of the simulation study is equal to the one used by AB for their Table III.

We focus on results for $k = 2, 4$, and 10. For each value of k, we consider three correlation matrices: $\Omega_{Neg}, \Omega_{Zero}$, and Ω_{Pos}. The matrix Ω_{Zero} equals I_k (that is, the identity matrix). The matrices Ω_{Neg} and Ω_{Pos} are Toeplitz matrices with correlations on the diagonals (as they go away from the main diagonal) given by the following. For $k = 2$: $\rho = -0.9$ for Ω_{Neg} and $\rho = 0.5$ for Ω_{Pos}. For $k = 4$: $\rho = (-0.9, 0.7, -0.5)$ for Ω_{Neg} and $\rho = (0.9, 0.7, 0.5)$ for Ω_{Pos}. For $k = 10$:

$\rho = (-0.9, 0.8, -0.7, 0.6, -0.5, 0.4, -0.3, 0.2, -0.1)$ for Ω_{Neg} and $\rho = (0.9, 0.8, 0.7, 0.6, 0.5, \ldots, 0.5)$ for Ω_{Pos}.

For $k = 2$, the set of μ vectors $\mathcal{M}_2(\Omega)$ for which asymptotic average power is computed includes seven elements: $\mathcal{M}_2(\Omega) = \{(\mu_1, 0), (\mu_2, 1), (\mu_3, 2), (\mu_4, 3), (\mu_5, 4), (\mu_6, 7), (\mu_7, \mu_7)\}$, where μ_j depends on Ω. For brevity, the values of μ_j in $\mathcal{M}_2(\Omega)$ and the sets $\mathcal{M}_k(\Omega)$ for $k = 4, 10$ are given in Section 7.1 of Andrews and Barwick (2012b). We point out, however, that we reverse the signs of the mean vectors used by AB, since in our framework the inequality signs are reversed in the null and alternative hypotheses.

To showcase the value, in terms of power properties, of incorporating information about which components of $\mathbb{E}_P[W_i]$ are "negative", we also include a one-step procedure which ignores such information. This one-step procedure simply uses $J_n^{-1}(1-\alpha, \lambda, \hat{P}_n)$ with the "least favorable" value of λ, i.e., $\lambda = 0$, as the critical value for the test statistic. Equivalently, it can be described as our two-step procedure using $\beta = 0$. Such an approach is expected to have higher power when all non-positive moments are equal to zero (or at least very close to zero) but is expected to have reduced power when some non-positive moments are far away from zero.

AB find that a bootstrap version of their test has better finite-sample size properties than a version based on asymptotic (normal) critical values. Therefore, we only implement bootstrap versions, both for the qlr, ad/t-Test/κAuto test and our two-step and one-step procedures. All bootstraps use $B = 499$ resamples; this is also the case for the first step of our two-step procedure.

The two-step procedure uses $\beta = 0.005$ for the construction of the confidence region in the first step. Using larger values of β leads to somewhat reduced average power in general. Lower values of β do not make a noticeable difference in terms of average power, but require a larger number of bootstrap resamples in the first step. (The reason is that the number of bootstrap samples needed to accurately estimate a β quantile is inversely related to β, for small values of β).

Unlike Andrews and Barwick (2012b), we do not consider any singular covariance matrices Ω. Therefore, the qlr, ad/t-Test/κAuto test as well as our two-step and one-step procedures use, for simplicity and reduced computational burden, the "unadjusted" quasi-likelihood ratio test statistic (7) rather than the "adjusted" version (8). For the scenarios that we consider, this does not make any difference.

4.1 Maximum Null Rejection Probabilities

Following AB, to ensure computational feasibility, empirical MNRPs are simulated as the maximum rejection probability over all μ vectors that are composed only of zero and $-\infty$ entries, containing at least one zero entry. So for dimension k, there are a total of $2^k - 1$ null vectors to consider. It is worth emphasizing, however, that it has not been proven that the maximum over these $2^k - 1$ null vectors equals the maximum over all μ vectors satisfying the null.

For each scenario, we use 10,000 repetitions to compute empirical MNRPs. The results are presented in the upper half of Table 1 and can be summarized as follows; from here on, we use the term AB-Rec to denote the recommended procedure of AB, i.e., the qlr, ad/t-Test/κAuto test.

- All procedures achieve a satisfactory performance.
- The empirical MNRP of the AB-Rec procedure is generally somewhat higher compared to the two-step and one-step procedures.
- The empirical MNRPs are somewhat higher when the distribution of the elements is heavy-tailed (i.e., t_3) or skewed (i.e., χ^2_3) versus standard normal.

4.2 Average Powers

Empirical average powers are computed over a set of m different alternative μ vectors, with $m = 7$ when $k = 2$, $m = 24$ when $k = 4$, and $m = 40$ when $k = 10$. For a fixed k, the specific set of μ vectors depends on the correlation matrix $\Omega \in \{\Omega_{Neg}, \Omega_{Zero}, \Omega_{Pos}\}$; see Andrews and Barwick (2012b, Subsection 7.2) for the details. For each scenario, we use 10,000 repetitions to compute empirical average powers when $k = 2$ and $k = 4$, and 5,000 repetitions to compute empirical average powers when $k = 10$. Unlike AB, we first report "raw" empirical average powers instead of size-corrected empirical average powers. If anything, this slightly favors the recommended procedure of AB, since our two-step and one-step procedures were seen to have (somewhat) lower empirical MNRPs in general. The results are presented in the lower half of Table 1 and can be summarized as follows.

- For every scenario, the AB-Rec procedure has the highest empirical average power and the one-step procedure has the lowest empirical average power. However, this does not mean that the AB-Rec procedure is uniformly more powerful than the other two procedures. For individual alternative μ vectors, even the one-step procedure can have higher empirical power than the AB-Rec procedure; for example, this happens when all non-positive moments are equal to zero.
- The two-step procedure generally picks up most of the difference in empirical average powers between AB-Rec and the one-step procedure; across the 27 scenarios, the average pickup is 74.1% and the median pickup is 76.4%. In particular, the relative improvement of the two-step procedure over the one-step procedure tends to be largest when it is needed most, i.e., when the differences between AB-Rec and the one-step procedure are the largest. Such cases correspond to Ω_{Neg}; across these 9 scenarios, the average pickup of the two-step procedure is 82.2% and the median pickup is 83.7%.

As mentioned before, reporting "raw" empirical average powers slightly favors the recommended procedure of AB, so we also compute "size-corrected" average powers for the two-step procedure. Because of the extremely high computational burden when $k = 10$, we are only able to do this for $k = 2$ and $k = 4$, however, as follows. For a given combination of $k \in \{2, 4\}$, $\Omega \in \{\Omega_{Neg}, \Omega_{Zero}, \Omega_{Pos}\}$, and $\text{Dist} \in \{N(0,1), t_3, \chi^2_3\}$, we vary the nominal level α for the two-step procedure, keeping $\beta = 0.005$ fixed, until the resulting MNRP matches that of the AB-Rec procedure with $\alpha = 0.05$. Denote the corresponding nominal level α for the two-step procedure by α_{sc}; for the 18 different combinations of (k, Ω, Dist) considered, we find that $\alpha_{\text{sc}} \in [0.051, 0.055]$. We then use α_{sc} to compute the "size-corrected" average empirical power for the given combination of (k, Ω, Dist). The results are presented in Table 2. The "fair" comparison is the one between AB-Rec and Two-Step$_{\text{sc}}$. It can be seen that the difference is always smaller than for the "unfair" comparison between AB-Rec and Two-Step.

4.3 Maximum Null Rejection Probabilities for a Large Number of Moment Inequalities

We finally turn attention to a case with a large number of inequalities, i.e., a case with $k > 10$, for which the procedures of AB are no longer available.

We feel that it is most informative to compute MNRPs. Since a comparison to AB-Rec (or any other of the procedures suggested by AB) is no longer possible, it is not clear what useful information could be taken away from computing empirical average powers.

As discussed before, computing MNRPs, in principle, involves the evaluation of $2^k - 1$ NRPs. Given current computational powers, this is infeasible for any value of k much larger than 10. However, for the special case of $\Omega = \Omega_{Zero}$, the problem is reduced to the evaluation of k NRPs only. This is because, under the identity covariance matrix, for a given number of zero entries, the position of these entries does not matter. So if there are p zero entries, say, one only has to evaluate a single NRP rather than $\binom{k}{p}$ NRPs; and without loss of generality, the corresponding single null vector can be chosen as $(0, \ldots, 0, -\infty, \ldots, -\infty)'$.

We use $k = 50$, which corresponds to roughly the limit of our computational capabilities. The sample sizes considered are $n = 100, 500$. It turns out that for $n = 100$, in many instances, the qlr test statistic cannot be computed because of numerical difficulties. We suspect that the reason is that for $(k = 50, n = 100)$, the sample covariance matrix is ill-conditioned; this problem is exacerbated in the bootstrap world where, in a given data set, there are always some repeated observations.

Therefore, in addition to the qlr test statistic, we also consider the following two alternative test statistics: first, the "modified method of moments" (MMM) test statistic T_n^{MMM} defined in (9) and second, the maximum test statistic $T_n^{\max}$ defined in (6).

For each scenario, we use 5,000 repetitions to compute empirical MNRPs. The results are presented in Tables 3 and can be summarized as follows.

- For $n = 100$, the results for the qlr test statistic are not available due to the numerical difficulties described above. The other two test statistics yield satisfactory performance throughout, though the one-step procedure is somewhat conservative when the distribution of the elements is heavy-tailed (i.e., t_3) or skewed (i.e., χ^2_3).
- For $n = 500$, both the two-step method and the one-step procedure yield satisfactory performance for all test statistics and all distributions of the elements considered.

A Appendix

In Appendix A.1, we establish a series of results that will be used in the proof of Theorem 2.1 in Appendix A.2. The proof of Theorem 3.1 is then provided in Appendix A.3.

A.1 Auxiliary Results

Lemma A.1. *Suppose μ_n is a sequence in $\mathbb{R}^k_-$ such that $\mu_n \to \mu$ with $\mu \in \overline{\mathbb{R}}^k_- = (\mathbb{R}_- \cup \{-\infty\})^k$. For $\tau \in \mathbb{R}^k$ and Γ a positive definite $k \times k$ real matrix, define*

$$f_n(\tau, \Gamma) = \inf_{t \in \mathbb{R}^k : t < -\mu_n} ||\tau - t||_\Gamma \ ,$$

where $||x||_\Gamma = (x'\Gamma x)^{\frac{1}{2}}$ for $x \in \mathbb{R}^k$. (Below, we may simply write $||x||$ for $||x||_{I_k}$.) Suppose $(\tau_n, \Gamma_n) \to (\tau, \Gamma)$, where Γ is positive definite. Then, $f_n(\tau_n, \Gamma_n) \to f(\tau, \Gamma)$, where

$$f(\tau, \Gamma) = \inf_{t \in \mathbb{R}^k : t < -\mu} ||\tau - t||_\Gamma \ .$$

Proof: We first argue that $f_n(\tau_n, \Gamma_n) - f_n(\tau, \Gamma) \to 0$. To see this, first note, by strict convexity and continuity of $||\Gamma^{\frac{1}{2}}(\tau - t)||$ as a function of $t \in \mathbb{R}^k$, that there exists $t^*_n \leq -\mu_n$ such that

$$\inf_{t \in \mathbb{R}^k : t < -\mu_n} ||\Gamma^{\frac{1}{2}}(\tau - t)|| = \min_{t \in \mathbb{R}^k : t \leq -\mu_n} ||\Gamma^{\frac{1}{2}}(\tau - t)|| = ||\Gamma^{\frac{1}{2}}(\tau - t^*_n)|| \ .$$

Next, since $0 \leq -\mu_n$, note that

$$||\Gamma^{\frac{1}{2}}(\tau - t^*_n)|| \leq ||\Gamma^{\frac{1}{2}}\tau|| \ . \tag{22}$$

Finally, observe that

$$\begin{aligned}
f_n(\tau_n, \Gamma_n) - f_n(\tau, \Gamma) &= \inf_{t \in \mathbb{R}^k : t < -\mu_n} ||\Gamma_n^{\frac{1}{2}}(\tau_n - t)|| - \inf_{t \in \mathbb{R}^k : t < -\mu_n} ||\Gamma^{\frac{1}{2}}(\tau - t)|| \\
&= \min_{t \in \mathbb{R}^k : t \leq -\mu_n} ||\Gamma_n^{\frac{1}{2}}(\tau_n - t)|| - \min_{t \in \mathbb{R}^k : t \leq -\mu_n} ||\Gamma^{\frac{1}{2}}(\tau - t)|| \\
&\leq ||\Gamma_n^{\frac{1}{2}}(\tau_n - t^*_n)|| - ||\Gamma^{\frac{1}{2}}(\tau - t^*_n)|| \\
&\leq ||\Gamma_n^{\frac{1}{2}}(\tau_n - t^*_n) - \Gamma^{\frac{1}{2}}(\tau - t^*_n)|| \\
&= ||\Gamma_n^{\frac{1}{2}}(\tau_n - \tau) + \Gamma_n^{\frac{1}{2}}(\tau - t^*_n) - \Gamma^{\frac{1}{2}}(\tau - t^*_n)|| \\
&= ||\Gamma_n^{\frac{1}{2}}(\tau_n - \tau) + \Gamma_n^{\frac{1}{2}}\Gamma^{-\frac{1}{2}}\Gamma^{\frac{1}{2}}(\tau - t^*_n) - \Gamma^{\frac{1}{2}}(\tau - t^*_n)|| \\
&\leq ||\Gamma_n^{\frac{1}{2}}(\tau_n - \tau)|| + ||\Gamma_n^{\frac{1}{2}}\Gamma^{-\frac{1}{2}} - I_k||_{op}||\Gamma^{\frac{1}{2}}(\tau - t^*_n)|| \\
&\leq ||\Gamma_n^{\frac{1}{2}}(\tau_n - \tau)|| + ||\Gamma_n^{\frac{1}{2}}\Gamma^{-\frac{1}{2}} - I_k||_{op}||\Gamma^{\frac{1}{2}}\tau|| \\
&\to 0 \ ,
\end{aligned}$$

where the first equality follows from the definition of the relevant norms, the second equality follows from strict convexity and continuity, the first inequality follows from the definition of t^*_n and the fact that $t_n^* \leq -\mu_n$, the second inequality follows from the reverse triangle inequality, the third and fourth equalities follow by inspection, the third inequality follows form the triangle inequality and the definition of the operator norm, the fourth inequality follows from (22), and the convergence to zero follows from the assumed convergences of τ_n and Γ_n.

Next, we argue that $f_n(\tau, \Gamma) \to f(\tau, \Gamma)$. For this purpose, it is useful to assume, without loss of generality, that $\mu_n = (\mu_n^{(1)}, \mu_n^{(2)})$ and $\mu = (\mu^{(1)}, \mu^{(2)})$, where all components of $\mu^{(1)}$ are finite and all components of $\mu^{(2)}$ are infinite. Define $\iota^{(1)}$ to be a vector of ones with the same length as $\mu^{(1)}$; define $\iota^{(2)}$ similarly. First note for $0 < \epsilon_n \to 0$ sufficiently slowly and n sufficiently large that

$$\begin{aligned} \inf_{t \in \mathbb{R}^k : t < -\mu_n} ||\tau - t||_\Gamma &\geq \inf_{t \in \mathbb{R}^k : t < -(\mu^{(1)}, \mu_n^{(2)}) + (\epsilon_n \iota^{(1)}, 0\iota^{(2)})} ||\tau - t||_\Gamma \\ &= \inf_{t \in \mathbb{R}^k : t < -(\mu^{(1)}, \mu_n^{(2)})} ||\tau - (\epsilon_n \iota^{(1)}, 0\iota^{(2)}) - t||_\Gamma \ . \end{aligned}$$

But, by identifying τ_n in the preceding paragraph with $\tau - (\epsilon_n \iota^{(1)}, 0\iota^{(2)})$ here, we see that the final expression equals

$$\inf_{t \in \mathbb{R}^k : t < -(\mu^{(1)}, \mu_n^{(2)})} ||\tau - t||_\Gamma + o(1) \ . \tag{23}$$

The same argument with $\epsilon < 0$ establishes that $\inf_{t \in \mathbb{R}^k : t < -\mu_n} ||\tau - t||_\Gamma$ in fact equals (23). To complete the argument, we argue that

$$\inf_{t \in \mathbb{R}^k : t < -(\mu^{(1)}, \mu_n^{(2)})} ||\tau - t||_\Gamma \to \inf_{t \in \mathbb{R}^k : t < -\mu} ||\tau - t||_\Gamma \ . \tag{24}$$

To establish this fact, given any subsequence n_k, consider a further subsequence n_{k_ℓ} such that $\mu_{n_{k_\ell}}^{(2)}$ is strictly increasing. By the monotone convergence theorem, we see that

$$\inf_{t \in \mathbb{R}^k : t < -(\mu^{(1)}, \mu_{n_{k_\ell}}^{(2)})} ||\tau - t||_\Gamma \to \inf_{t \in \mathbb{R}^k : t < -\mu} ||\tau - t||_\Gamma \ .$$

Hence, (24) holds. ■

Lemma A.2. *Let $W_i, i = 1, \ldots, n$, be an i.i.d. sequence of random variables with distribution $P \in \mathbf{P}$ on $\mathbb{R}^k$, where $\mathbf{P}$ satisfies (16). Then, $M_n(1-\beta)$ defined by (4) satisfies*

$$\liminf_{n \to \infty} \inf_{P \in \mathbf{P}} P\{\mu(P) \in M_n(1-\beta)\} \geq 1 - \beta \ . \tag{25}$$

PROOF: Follows immediately from Theorem 3.7 in Romano and Shaikh (2012). ■

Lemma A.3. *Consider a sequence $\{P_n \in \mathbf{P} : n \geq 1\}$ where $\mathbf{P}$ is a set of distributions on $\mathbb{R}^k$ satisfying (16). Let $W_i, , i = 1, \ldots, n$, be an i.i.d. sequence of random variables with distribution P_n. Suppose*

$$\frac{\sqrt{n}\mu_j(P_n)}{\sigma_j(P_n)} \to -\infty \ ,$$

for all $1 \leq j \leq k$. Then,

$$P_n\{M_n(1-\beta) \subseteq \mathbb{R}^k_-\} \to 1 \ .$$

Proof: Note that we may write $M_n(1-\beta)$ as the set of all $\mu \in \mathbb{R}^k$ such that

$$\mu_j \leq \frac{\sigma_j(P_n)}{\sqrt{n}}\left[\frac{\sqrt{n}\big(\bar{X}_{j,n}-\mu_j(P_n)\big)}{\sigma_j(P_n)}+\frac{\sqrt{n}\mu_j(P_n)}{\sigma_j(P_n)}+\frac{K_n^{-1}(1-\beta,\hat{P}_n)}{\sigma_j(P_n)}S_{j,n}\right]$$

for all $1 \leq j \leq k$. From Lemma 11.4.1 of Lehmann and Romano (2005), we see that

$$\frac{\sqrt{n}\big(\bar{W}_{j,n}-\mu_j(P_n)\big)}{\sigma_j(P_n)} = O_{P_n}(1)\ .$$

By assumption,

$$\frac{\sqrt{n}\mu_j(P_n)}{\sigma_j(P_n)} \to -\infty\ .$$

From Lemma 4.8 in Romano and Shaikh (2012), we see that

$$\frac{S_{j,n}}{\sigma_j(P_n)} \stackrel{P_n}{\to} 1\ .$$

Finally, note that

$$K_n^{-1}(1-\beta,\hat{P}_n) = O_{P_n}(1)$$

because, using the Bonferroni inequality, it is asymptotically bounded above by $\Phi^{-1}(1-\beta/k)$, from which the desired result follows. ■

Lemma A.4. *Let* $\mathbf{P}'$ *be the set of all distributions on* $\mathbb{R}^k$ *and let* $\mathbf{P}$ *be a set of distributions on* $\mathbb{R}^k$ *satisfying* (16). *For* $(P,Q) \in \mathbf{P}' \times \mathbf{P}$, *define*

$$\rho(Q,P) = \max\left\{\max_{1\leq j\leq k}\left\{\int_0^\infty |r_j(\lambda,Q)-r_j(\lambda,P)|\exp(-\lambda)d\lambda\right\}, \max_{1\leq j\leq k}\left|\frac{\sigma_j(P)}{\sigma_j(Q)}-1\right|, ||\Omega(Q)-\Omega(P)||\right\}\ ,$$

where

$$r_j(\lambda,P) = \mathbb{E}_P\left[\left(\frac{X_j-\mu_j(P)}{\sigma_j(P)}\right)^2 \mathbb{1}\left\{\left|\frac{X_j-\mu_j(P)}{\sigma_j(P)}\right| > \lambda\right\}\right]\ , \tag{26}$$

and the norm $||\cdot||$ *is the component-wise maximum of the absolute value of all elements. Let* $\{Q_n \in \mathbf{P}' : n \geq 1\}$ *and* $\{P_n \in \mathbf{P} : n \geq 1\}$ *be such that* $\rho(P_n,Q_n) \to 0$ *and for some* $\emptyset \neq I \subseteq \{1,\ldots k\}$,

$$\frac{\sqrt{n}\lambda_{j,n}}{\sigma_j(P_n)} \to -\delta_j \quad \textit{for all } j \in I \textit{ and some } \delta_j \geq 0$$

and

$$\frac{\sqrt{n}\lambda_{j,n}}{\sigma_j(P_n)} \to -\infty \quad \textit{for all } j \notin I\ .$$

Then, for T_n *given by* (6), (8) *or* (9), *we have*

$$\limsup_{n\to\infty}\sup_{x\in\mathbb{R}^k}\big|J_n(x,\lambda_n,P_n)-J_n(x,\lambda_n,Q_n)\big| = 0\ . \tag{27}$$

Proof: Consider first the case where T_n is given by (6). Note that

$$\frac{\sqrt{n}\lambda_{j,n}}{S_{j,n}} = \frac{\sigma_j(P_n)}{S_{j,n}} \frac{\sqrt{n}\lambda_{j,n}}{\sigma_j(P_n)} .$$

From Lemma 4.8 in Romano and Shaikh (2012), we see that

$$\frac{S_{j,n}}{\sigma_j(P_n)} \stackrel{P_n}{\to} 1 .$$

Hence,

$$\frac{\sqrt{n}\lambda_{j,n}}{S_{j,n}} \stackrel{P_n}{\to} -\delta_j \quad \text{for all } j \in I \tag{28}$$

and

$$\frac{\sqrt{n}\lambda_{j,n}}{S_{j,n}} \stackrel{P_n}{\to} -\infty \quad \text{for all } j \notin I . \tag{29}$$

It follows that

$$\max_{1 \le j \le k} \left(\frac{\sqrt{n}(\bar{W}_{j,n} - \mu_j(P_n))}{S_{j,n}} + \frac{\sqrt{n}\lambda_{j,n}}{S_{j,n}} \right) = \max_{j \in I} \left(\frac{\sqrt{n}(\bar{W}_{j,n} - \mu_j(P_n))}{S_{j,n}} + \frac{\sqrt{n}\lambda_{j,n}}{S_{j,n}} \right) + o_{P_n}(1) . \tag{30}$$

Next, we argue that

$$\max_{1 \le j \le k} \left(\frac{\sqrt{n}(\bar{W}_{j,n} - \mu_j(Q_n))}{S_{j,n}} + \frac{\sqrt{n}\lambda_{j,n}}{S_{j,n}} \right) = \max_{j \in I} \left(\frac{\sqrt{n}(\bar{W}_{j,n} - \mu_j(Q_n))}{S_{j,n}} + \frac{\sqrt{n}\lambda_{j,n}}{S_{j,n}} \right) + o_{Q_n}(1) . \tag{31}$$

For this purpose, it suffices to show that the convergences in (28) and (29) also hold with P_n replaced by Q_n. To see this, first note that by arguing as in the proof of Lemma 4.11 in Romano and Shaikh (2012) we have that

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} r_j(\lambda, Q_n) = 0 .$$

The convergence $\rho(P_n, Q_n) \to 0$ implies further that

$$\frac{\sigma_j(P_n)}{\sigma_j(Q_n)} \to 1 \quad \text{for all } 1 \le j \le k .$$

Since

$$\frac{\sqrt{n}\lambda_{j,n}}{S_{j,n}} = \frac{\sigma_j(Q_n)}{S_{j,n}} \frac{\sigma_j(P_n)}{\sigma_j(Q_n)} \frac{\sqrt{n}\lambda_{j,n}}{\sigma_j(P_n)} ,$$

the desired conclusion follows. Finally, (27) now follows from (30) and (31) and by arguing as in the proof of Lemma 4.11 in Romano and Shaikh (2012).

Now consider the case where T_n is given by (8). Note that

$$T_n^{\text{qlr,ad}} = \inf_{t \in \mathbb{R}^k : t < -\sqrt{n} D^{-1}(P_n)\lambda_n} \tilde{Z}_n(t)' \tilde{\Omega}_n D^2(P_n) S_n^{-2} \tilde{Z}_n(t) ,$$

where

$$\tilde{Z}_n(t) = \left(\frac{\sqrt{n}(\bar{W}_{1,n}-\mu_1(P_n))}{\sigma_1(P_n)} - t_1, \ldots, \frac{\sqrt{n}(\bar{W}_{k,n}-\mu_k(P_n))}{\sigma_k(P_n)} - t_k\right)'$$
$$D^2(P_n) = \text{diag}\left(\sigma_1^2(P_n), \ldots, \sigma_k^2(P_n)\right) .$$

Now suppose by way of contradiction that (27) fails. It follows that there exists a subsequence n_k along which the left-hand side of (27) converges to a non-zero constant and

$$\Omega(P_{n_k}) \to \Omega^* \text{ , as well as} \tag{32}$$

$$\left(\frac{\bar{W}_{1,n_k}-\mu_1(P_{n_k}))}{\sigma_1(P_{n_k})}, \ldots, \frac{\bar{W}_{k,n_k}-\mu_k(P_{n_k}))}{\sigma_k(P_{n_k})}\right)' \stackrel{d}{\to} Z \sim N(0,\Omega^*) \text{ under } P_{n_k} . \tag{33}$$

Since

$$D^2(P_{n_k})S_{n_k}^{-2} \to I_k ,$$

we have further that

$$\tilde{\Omega}_{n_k} D^2(P_{n_k})S_{n_k}^{-2} \stackrel{P_{n_k}}{\to} \max\{\epsilon - \det(\Omega^*), 0\}I_k + \Omega^* = \bar{\Omega} . \tag{34}$$

Note that along such a subsequence n_k we also have that

$$\frac{\sqrt{n_k}\lambda_{j,n_k}}{\sigma_j(P_{n_k})} \to -\delta_j \quad \text{for all } j \in I \tag{35}$$

and

$$\frac{\sqrt{n_k}\lambda_{j,n_k}}{\sigma_j(P_{n_k})} \to -\infty \quad \text{for all } j \notin I . \tag{36}$$

Hence, by Lemma A.1 and the extended continuous mapping theorem (van der Vaart and Wellner, 1996; Theorem 1.11.1), we have that

$$T_{n_k}^{\text{qlr,ad}} \stackrel{d}{\to} \inf_{t\in\mathbb{R}^k : t<-\delta} (Z-t)'\bar{\Omega}^{-1}(Z-t) \text{ under } P_{n_k} . \tag{37}$$

Note that a similar result under slightly stronger assumptions could be established using, for example, Lemma S.1 in Bugni et al. (2012). Moreover, by Chow and Teicher (1978, Lemma 3, p.260), we have that

$$\sup_{x\in\mathbf{R}} \left| P_{n_k}\{T_{n_k}^{\text{qlr,ad}} \le x\} - P\Big\{\inf_{t\in\mathbb{R}^k : t<-\delta} (Z-t)'\bar{\Omega}^{-1}(Z-t) \le x\Big\}\right| \to 0 ,$$

since the distribution of $\inf_{t\in\mathbb{R}^k:t<-\delta}(Z-t)'\bar{\Omega}^{-1}(Z-t)$ is continuous everywhere except possibly at zero and

$$P_{n_k}\{T_{n_k}^{\text{qlr,ad}} \le 0\} \to P\{Z \le -\delta\} = P\Big\{\inf_{t\in\mathbb{R}^k : t<-\delta} (Z-t)'\bar{\Omega}^{-1}(Z-t) \le 0\Big\} .$$

Next, note that by arguing as above it follows from the assumed convergence $\rho(P_{n_k}, Q_{n_k}) \to 0$ that (32) – (36) all hold when P_{n_k} is replaced by Q_{n_k}. Hence, by the triangle inequality, we see that along n_k, the lefthand-side of (27) must converge to zero, from which the desired result follows.

Finally, consider the test statistic (9), for which the argument is easier. For example, the above argument for (8) can be used with $\tilde{\Omega}_n$ replaced by the identity, so that the convergence (37) holds with $\bar{\Omega}$ replaced by the identity. ■

Lemma A.5. *Consider a sequence $\{P_n \in \mathbf{P} : n \geq 1\}$ where $\mathbf{P}$ is a set of distributions on $\mathbb{R}^k$ satisfying (16). Let $W_{n,i}, i = 1, \ldots, n$, be an i.i.d. sequence of random variables with distribution P_n. Suppose that for some $\emptyset \neq I \subseteq \{1, \ldots k\}$,*

$$\frac{\sqrt{n}\mu_j(P_n)}{\sigma_j(P_n)} \to -\delta_j \quad \text{for all } j \in I \text{ and some } \delta_j \geq 0$$

and

$$\frac{\sqrt{n}\mu_j(P_n)}{\sigma_j(P_n)} \to -\infty \quad \text{for all } j \notin I \ .$$

Then,

$$P_n\{T_n > J_n^{-1}(1 - \alpha + \beta, \mu(P_n), \hat{P}_n)\} \to \alpha - \beta \ .$$

PROOF: Let $\mathbf{P}'$ and $\rho(P, Q)$ be defined as in Lemma A.4. Trivially,

$$P_n\{\hat{P}_n \in \mathbf{P}'\} \to 1 \ .$$

From Lemma 4.8 in Romano and Shaikh (2012), we see that

$$\max_{1 \leq j \leq k} \left| \frac{S_{j,n}}{\sigma_j(P_n)} - 1 \right| \stackrel{P_n}{\to} 0 \ .$$

From Lemma 4.9 in Romano and Shaikh (2012), we see that

$$||\Omega(\hat{P}_n) - \Omega(P_n)|| \stackrel{P_n}{\to} 0 \ .$$

It follows from Lemma 4.12 in Romano and Shaikh (2012) that

$$\rho(\hat{P}_n, P_n) \stackrel{P_n}{\to} 0 \ .$$

The desired result now follows by applying Lemma A.4 with $\lambda_n = \mu(P_n)$ and Theorem 2.4 in Romano and Shaikh (2012). ■

A.2 Proof of Theorem 2.1

Suppose by way of contradiction that (3) fails. It follows that there exist a subsequence n_k and $\eta > \alpha$ such that

$$\mathbb{E}_{P_{n_k}}[\phi_{n_k}] \to \eta \ . \tag{38}$$

There are two cases to consider.

First, consider the case where there exists a further subsequence (which, by an abuse of notation, we continue to denote by n_k) such that

$$\frac{\sqrt{n}_k \mu_j(P_{n_k})}{\sigma_j(P_{n_k})} \to -\infty$$

for all $1 \leq j \leq k$. Then, by Lemma A.3, we see that

$$P_{n_k}\{M_{n_k}(1-\beta) \subseteq \mathbb{R}^k_-\} \to 1 \ .$$

Hence,

$$\mathbb{E}_{P_{n_k}}[\phi_{n_k}] \to 0 \ ,$$

contradicting (38).

Second, consider the case where there exists a further subsequence (which, by an abuse of notation, we continue to denote by n_k) and $\emptyset \neq I \subseteq \{1, \dots k\}$ such that

$$\frac{\sqrt{n}_k \mu_j(P_{n_k})}{\sigma_j(P_{n_k})} \to -\delta_j \quad \text{for all } j \in I \text{ and some } \delta_j \geq 0$$

and

$$\frac{\sqrt{n_k} \mu_j(P_{n_k})}{\sigma_j(P_{n_k})} \to -\infty \quad \text{for all } j \notin I \ .$$

Next, recall the definition of $\hat{c}_n(1-\alpha+\beta)$ in (14) and note that

$$\begin{aligned}\mathbb{E}_{P_{n_k}}[\phi_{n_k}] &\leq P_{n_k}\{T_{n_k} > \hat{c}_{n_k}(1-\alpha+\beta)\} \\ &\leq P_{n_k}\{T_{n_k} > J^{-1}_{n_k}(1-\alpha+\beta, \mu(P_{n_k}), \hat{P}_{n_k})\} + P_{n_k}\{\mu(P_{n_k}) \notin M_{n_k}(1-\beta)\} \ .\end{aligned}$$

Then, by Lemmas A.2 and A.5, we have that

$$\limsup_{k\to\infty} \mathbb{E}_{P_{n_k}}[\phi_{n_k}] \leq \alpha \ ,$$

contradicting (38). ■

A.3 Proof of Theorem 3.1

Follows immediately from Theorem 2.1 by identifying the distribution of $g(X_i, \theta)$ under $P \in \mathbf{P}$ and $\theta \in \Theta_0(P)$ in the present context with the distribution of W_i under P in Theorem 2.1. ■

References

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68:399–405.

Andrews, D. W. K. and Barwick, P. J. (2012a). Inference for parameters defined by moment inequalities: A recommended moment selection procedure. *Econometrica*, 80(6):2805–2826.

Andrews, D. W. K. and Barwick, P. J. (2012b). Inference for parameters defined by moment inequalities: A recommended moment selection procedure. *Econometrica Supplemental Material*, 80(6). Available at `http://www.econometricsociety.org/`.

Andrews, D. W. K. and Guggenberger, P. (2009). Validity of subsampling and 'plug-in asymptotic' inference for parameters defined by moment inequalities. *Econometric Theory*, 25(3):669–709.

Andrews, D. W. K. and Guggenberger, P. (2010). Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory*, 26(2):426–468.

Andrews, D. W. K. and Soares, G. (2010). Inference for parameters defined by momoment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157.

Bajari, P., Benkard, C., and Levin, J. (2007). Estimating dynamic models of imperfect competition. *Econometrica*, 75(5):1331–1370.

Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016.

Bugni, F. (2011). A comparison of inferential methods in partially identified models in terms of error in the coverage probability. Working paper, Department of Economics, Duke University.

Bugni, F. A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica*, 78(2):735–753.

Bugni, F. A., Canay, I. A., and Guggenberger, P. (2012). Distortions of asymptotic confidence size in locally misspecified moment inequality models. *Econometrica*, 80(4):1741–1768.

Canay, I. A. (2010). EL inference for partially identified models: large deviations optimality and bootstrap validity. *Journal of Econometrics*, 156(2):408–425.

Chernozhukov, V., Hong, H., and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75(5):1243–1284.

Chow, Y. S. and Teicher, H. (1978). *Probability Theory: Independence, Interchangeability and Martingales*. Springer, New York.

Ciliberto, F. and Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6):1791–1828.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economics Statistics*, 23:365–380.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, third edition.

Loh, W. (1985). A new method for testing separate families of hypotheses. *Journal of the American Statistical Association*, 80:362–368.

McCloskey, A. (2012). Bonferroni-based size correction for nonstandard testing problems. Working paper, Department of Economics, Brown University.

Romano, J. P. and Shaikh, A. M. (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference – Special Issue in Honor of Ted Anderson*, 138(9):2786–2807.

Romano, J. P. and Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. *Econometrica*, 78(1):169–211.

Romano, J. P. and Shaikh, A. M. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *Annals of Statistics*, 40(6):2798–2822.

Romano, J. P. and Wolf, M. (2000). Finite sample nonparametric inference and large sample efficiency. *Annals of Statistics*, 28(3):756–778.

Silvapulle, M. J. (1996). A test in the presence of nuisance parameters. *Journal of the American Statistical Association*, 91(436):1690–1693.

Stock, J. and Staiger, D. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

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

Table 1: Empirical maximum null rejection probabilities (MNRPs), upper half, and empirical average powers, lower half, of the AB-recommended procedure, the two-step procedure, and the one-step procedure. The nominal level is $\alpha = 5\%$ and the sample size is $n = 100$. All results are based on 10,000 repetitions when $k = 2, 4$ and on 5,000 repetitions when $k = 10$.

			$k=2$			$k=4$			$k=10$		
Test	Dist	H_0/H_1	Ω_{Neg}	Ω_{Zero}	Ω_{Pos}	Ω_{Neg}	Ω_{Zero}	Ω_{Pos}	Ω_{Neg}	Ω_{Zero}	Ω_{Pos}
AB-Rec	$N(0,1)$	H_0	5.3	5.1	4.9	5.3	5.0	5.1	5.8	5.9	5.6
Two-Step	$N(0,1)$	H_0	5.0	4.8	4.5	5.1	4.9	5.0	5.3	5.2	5.4
One-Step	$N(0,1)$	H_0	5.2	5.1	4.9	4.9	5.0	5.1	5.2	4.9	5.3
AB-Rec	t_3	H_0	6.2	6.2	5.9	5.7	5.9	5.7	5.4	5.5	5.3
Two-Step	t_3	H_0	5.6	5.7	5.6	5.3	5.7	5.4	5.7	5.6	5.6
One-Step	t_3	H_0	5.2	6.1	5.7	4.7	5.3	5.7	5.3	5.2	5.7
AB-Rec	χ^2_3	H_0	5.2	4.9	5.1	5.3	4.8	4.9	5.8	5.9	6.0
Two-Step	χ^2_3	H_0	4.8	4.4	4.8	5.1	4.7	4.8	5.6	5.3	5.7
One-Step	χ^2_3	H_0	4.6	4.9	5.1	4.9	5.0	5.0	5.3	4.9	5.5
AB-Rec	$N(0,1)$	H_1	64.1	68.1	71.4	59.1	66.6	77.5	54.7	63.6	78.9
Two-Step	$N(0,1)$	H_1	62.0	65.1	66.4	56.1	60.6	74.4	51.0	54.8	75.6
One-Step	$N(0,1)$	H_1	52.7	61.1	64.2	41.3	50.4	72.6	23.9	32.6	68.4
AB-Rec	t_3	H_1	68.1	72.4	75.2	63.9	71.5	79.5	58.9	68.2	80.4
Two-Step	t_3	H_1	66.0	69.1	71.0	61.1	66.1	76.6	54.9	58.9	77.4
One-Step	t_3	H_1	61.7	66.2	68.8	46.7	57.2	74.9	27.6	37.7	71.5
AB-Rec	χ^2_3	H_1	69.3	76.4	77.9	63.1	74.5	82.4	57.8	69.8	82.6
Two-Step	χ^2_3	H_1	67.6	73.7	74.3	61.0	70.8	80.1	55.5	63.7	80.7
One-Step	χ^2_3	H_1	63.7	70.1	71.7	46.9	59.5	77.9	26.1	37.2	73.5

Table 2: Empirical average powers of the AB-recommended procedure and the two-step procedure and empirical "size-corrected" average powers of the two-step procedure. The nominal level is $\alpha = 5\%$ and the sample size is $n = 100$. Empirical (size-corrected) average powers are based on 10,000 repetitions.

			$k=2$			$k=4$		
Test	Dist	H_0/H_1	Ω_{Neg}	Ω_{Zero}	Ω_{Pos}	Ω_{Neg}	Ω_{Zero}	Ω_{Pos}
AB-Rec	$N(0,1)$	H_1	64.1	68.1	71.4	59.1	66.6	77.5
Two-Step$_{\text{sc}}$	$N(0,1)$	H_1	63.3	66.3	67.8	56.7	62.1	75.2
Two-Step	$N(0,1)$	H_1	62.0	65.1	66.4	56.1	60.6	74.4
AB-Rec	t_3	H_1	68.1	72.4	75.2	63.9	71.5	79.5
Two-Step$_{\text{sc}}$	t_3	H_1	67.5	70.2	72.4	61.7	67.0	77.3
Two-Step	t_3	H_1	66.0	69.1	71.0	61.1	66.1	76.6
AB-Rec	χ^2_3	H_1	69.3	76.4	77.9	63.1	74.5	82.4
Two-Step$_{\text{sc}}$	χ^2_3	H_1	69.0	74.8	75.6	61.8	71.8	80.6
Two-Step	χ^2_3	H_1	67.6	73.7	74.3	61.0	70.8	80.1

Table 3: Empirical maximum null rejection probabilities of the two-step procedure and the one-step procedure based on various test statistics. The nominal level is $\alpha = 5\%$ and the covariance matrix is Ω_{Zero}. All results are based on 5,000 repetitions.

			$k=50, n=100$			$k=50, n=500$		
Test	Dist	H_0/H_1	T_n^{qlr}	T_n^{MMM}	$T_n^{\max}$	T_n^{qlr}	T_n^{MMM}	$T_n^{\max}$
Two-Step	$N(0,1)$	H_0	NA	4.9	5.1	4.9	4.8	5.1
One-Step	$N(0,1)$	H_0	NA	4.5	4.9	5.2	5.1	5.2
Two-Step	t_3	H_0	NA	4.3	4.4	4.4	4.7	4.9
One-Step	t_3	H_0	NA	2.9	2.1	4.7	4.5	4.0
Two-Step	χ^2_3	H_0	NA	4.5	4.7	5.2	5.2	5.1
One-Step	χ^2_3	H_0	NA	3.0	4.3	4.9	5.0	5.2

Supplement to
“A Practical Two-Step Method for Testing Moment Inequalities”

Joseph P. Romano *
Departments of Statistics and Economics
Stanford University
romano@stanford.edu

Azeem M. Shaikh †
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Michael Wolf
Department of Economics
University of Zurich
michael.wolf@econ.uzh.ch

April 2014

Abstract

This document provides additional results for the authors’ paper “A Practical Two-Step Method for Testing Moment Inequalities”.

*Research supported by NSF Grant DMS-0707085.
†Research supported by NSF Grant DMS-1227091 and the Alfred P. Sloan Foundation.

S.1 The Gaussian Problem

In this section, we assume that $W = (W_1, \ldots, W_k)' \sim P \in \mathbf{P} = \{N(\mu, \Sigma) : \mu \in \mathbb{R}^k\}$ for a known covariance matrix Σ. In this setting, we may equivalently describe the problem of testing (1) as the problem of testing

$$H_0 : \mu \in \Omega_0 \text{ versus } H_1 : \mu \in \Omega_1 \ , \tag{S.1}$$

where

$$\Omega_0 = \{\mu : \ \mu_j \leq 0 \text{ for } 1 \leq j \leq k\} \tag{S.2}$$

and $\Omega_1 = \mathbb{R}^k \setminus \Omega_0$. Here, it is possible to obtain some exact results, so we focus on tests $\phi_n = \phi_n(W_1, \ldots, W_n)$ of (S.1) that satisfy

$$\sup_{\mu \in \Omega_0} \mathbb{E}_P[\phi_n] \leq \alpha \tag{S.3}$$

for some pre-specified value of $\alpha \in (0, 1)$ rather than (3). In Section S.1 below, we first establish an upper bound on the power function of any test of (S.1) that satisfies (S.3) by deriving the most powerful test against any fixed alternative. We then describe our two-step procedure for testing (S.1) in Section S.2. Proofs of all results can be found in the Appendix.

Before proceeding, note that by sufficiency we may assume without loss of generality that $n = 1$. Hence, the data consists of a single random variable W distributed according to the multivariate Gaussian distribution with unknown mean vector $\mu \in \mathbb{R}^k$ and known covariance matrix Σ. For $1 \leq j \leq k$, we will denote by W_j the jth component of W and by μ_j the jth component of μ. Note further that, because Σ is assumed known, we may assume without loss of generality that its diagonal consists of ones; otherwise, we can simply replace W_j by W_j divided by its standard deviation.

S.1 Power Envelope

In this subsection only, we assume further that Σ is invertible.

Below we calculate the most powerful (MP) test of $\mu \in \Omega_0$ satisfying (S.3) against a fixed alternative $\mu = a$, where $a \in \Omega_1$. The power of such a test, as a function of a, provides an upper bound on the power function of any test of (S.1) satisfying (S.3) and is, therefore, referred to as the power envelope function. In Andrews and Barwick (2012a,b), numerical evidence is given to justify their conjecture of how to calculate the MP test of $\mu \in \Omega_0$ satisfying (S.3) against $\mu = a$ and hence how to calculate the power envelope function. Theorem S.1.1 below verifies the claim made by Andrews and Barwick (2012a). Note that the power of the MP test of $\mu \in \Omega_0$ satisfying (S.3) against $\mu = a$ depends on a through its "distance" from Ω_0 in terms of the Mahanolobis metric $d(x, y) = \sqrt{(x - y)'\Sigma^{-1}(x - y)}$, i.e.,

$$\inf_{\mu \in \Omega_0} \sqrt{\{(\mu - a)'\Sigma^{-1}(\mu - a)\}} \ . \tag{S.4}$$

Theorem S.1.1. *Let W be multivariate normal with unknown mean vector μ and known covariance matrix Σ. For testing $\mu \in \Omega_0$ against the fixed alternative $\mu = a$, where $a \in \Omega_1$, the MP test satisfying (S.3) rejects for large values of $T = W'\Sigma^{-1}(a - \bar{\mu})$, where*

$$\bar{\mu} = \underset{\mu \in \Omega_0}{\operatorname{argmin}} (\mu - a)'\Sigma^{-1}(\mu - a) \ .$$

In fact, the distribution which puts mass one at the point $\bar{\mu}$ is least favorable, and the critical value at level α can be determined so that

$$P_{\bar{\mu}}\{T > c_{1-\alpha}\} = \alpha \ .$$

Under $\mu = \bar{\mu}$,

$$\begin{aligned} \mathbb{E}[T] &= \bar{\mu}'\Sigma^{-1}(a - \bar{\mu}) \\ Var[T] &= (\bar{\mu} - a)'\Sigma^{-1}(\bar{\mu} - a) \ , \end{aligned}$$

so

$$c_{1-\alpha} = \bar{\mu}'\Sigma^{-1}(a - \bar{\mu}) + z_{1-\alpha}\sqrt{(\bar{\mu} - a)'\Sigma^{-1}(\bar{\mu} - a)} \ ,$$

where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard normal distribution. Moreover, the power of this test is given by

$$1 - \Phi\left(z_{1-\alpha} - \sqrt{(\bar{\mu} - a)'\Sigma^{-1}(\bar{\mu} - a)}\right) \ ,$$

where $\Phi(\cdot)$ denotes the standard normal c.d.f.

Since the most powerful tests vary as a function of the vector a, it follows that there is no uniformly most powerful test. Furthermore, as argued in Lehmann (1952), the only unbiased test is the trivial test whose power function is constant and equal to α. Invariance considerations do not appear to lead to any useful simplification of the problem either; also see Andrews (2012) for some negative results concerning similarity.

Remark S.1.1. Note that $T = W'\Sigma^{-1}(a-\bar{\mu})$ in Theorem S.1.1 is a linear combination $\sum_{1 \le j \le k} c_j W_j$ of the $W_1, \ldots, W_k$. Even if all components of a are positive, depending on Σ, $\bar{\mu}$ may not equal zero. One might, therefore, suspect that the test described in Theorem S.1.1 does not satisfy (S.3). However, the proof of the theorem shows that if $\bar{\mu}$ has any components that are negative, then the corresponding coefficient of W_j in T must be zero; components of $\bar{\mu}$ that are zero have corresponding coefficient of W_j in T that are nonnegative. ■

S.2 A Two-Step Procedure

There are, of course, many ways in which to construct a test of (S.1) that controls size at level α. For instance, given any test statistic $T = T(W_1, \ldots, W_k)$ that is nondecreasing in each of its arguments, we may consider a test that rejects H_0 for large values of T. Note that, for any given fixed critical value c, $P_\mu\{T(W_1,\ldots,W_k) > c\}$ is a nondecreasing function of μ. Therefore, if $c = c_{1-\alpha}$ is chosen to satisfy

$$P_0\big\{T(W_1,\ldots,W_k) > c_{1-\alpha}\big\} \le \alpha \ ,$$

then the test that rejects H_0 when $T > c_{1-\alpha}$ is a level α test. A reasonable choice of test statistic T is the likelihood ratio statistic, which is given by

$$T = \inf_{\mu \in \Omega_0} \{(W-\mu)'\Sigma^{-1}(W-\mu)\} \ . \tag{S.5}$$

By analogy with (S.4) and Theorem S.1.1, rejecting for large values of the "distance" of W to Ω_0 is intuitively appealing. It is easy to see that such a test statistic T is nondecreasing in each of its arguments.

A second choice of monotone test statistic is the "modified method of moments" test statistic

$$T = \sum_{j=1}^{k} W_j^2 \cdot \mathbb{1}\{W_j > 0\} \ .$$

A further choice of monotone test statistic is the maximal order statistic $T = \max\{W_1,\ldots,W_k\}$. For any given choice of monotone test statistic, a critical value $c_{1-\alpha}$ may be determined as the $1-\alpha$ quantile of the distribution of T when $(W_1,\ldots,W_k)'$ is multivariate normal with mean 0 and covariance matrix Σ. Unfortunately, as k increases, so does the critical value, which can make it difficult to have any reasonable power against alternatives. The main idea of our procedure, as well as that of Andrews and Barwick (2012a), is to essentially remove from consideration those μ_j that are "negative." If we can eliminate such μ_j from consideration, then we may use a smaller critical value with the hopes of increased power against alternatives.

Using this reasoning as a motivation, we may use a confidence region to help determine which μ_j are "negative." To this end, let $M(1-\beta)$ denote an upper confidence rectangle for all the μ_j simultaneously at level $1-\beta$. Specifically, let

$$\begin{aligned} M(1-\beta) &= \big\{\mu \in \mathbb{R}^k : \max_{1\le j\le k}(\mu_j - W_j) \le K^{-1}(1-\beta)\big\} \\ &= \big\{\mu \in \mathbb{R}^k : \mu_j \le W_j + K^{-1}(1-\beta) \ \text{ for all } 1 \le j \le k\big\} \ , \end{aligned} \tag{S.6}$$

where $K^{-1}(1-\beta)$ is the $1-\beta$ quantile of the distribution

$$K(x) = P_\mu\Big\{\max_{1\le j\le k}(\mu_j - W_j) \le x\Big\} \ .$$

Note that $K(\cdot)$ depends only on the dimension k and the underlying covariance matrix Σ. In particular, it does not depend on the μ_j, so it can be computed under the assumption that all $\mu_j = 0$. By construction, we have for any $\mu \in \mathbb{R}^k$ that

$$P_\mu\{\mu \in M(1-\beta)\} = 1-\beta \ .$$

The idea is that with probability at least $1-\beta$, we may assume that under the null hypothesis, μ in fact will lie in $\Omega_0 \cap M(1-\beta)$ rather than just Ω_0. Instead of computing the critical value under $\mu = 0$, the largest value of μ in Ω_0, we may, therefore, compute the critical value under $\tilde{\mu}$, the "largest" value of μ in the (data-dependent) set $\Omega_0 \cap M(1-\beta)$. It is straightforward to determine $\tilde{\mu}$ explicitly. In particular, $\tilde{\mu}$ has jth component equal to

$$\tilde{\mu}_j = \min\{W_j + K^{-1}(1-\beta), 0\} \ . \tag{S.7}$$

But, to account for the fact that μ may not lie in $M(1-\beta)$ with probability at most β, we reject H_0 when $T(W_1, \ldots, W_k)$ exceeds the $1-\alpha+\beta$ quantile of the distribution of T under $\tilde{\mu}$ rather than the $1-\alpha$ quantile of the distribution of T under $\tilde{\mu}$. Such an adjustment is in the same spirit as the "size correction factor" in Andrews and Barwick (2012a), but requires no computation to determine; see Remark S.1.5 for further discussion. The following theorem establishes that this test of (S.1) satisfies (S.3).

Theorem S.1.2. *Let $T(W_1, \ldots, W_k)$ denote any test statistic that is nondecreasing in each of its arguments. For $\mu \in \mathbb{R}^k$ and $\gamma \in (0,1)$, define*

$$b(\gamma, \mu) = \inf\{x \in \mathbb{R} : P_\mu\{T(W_1, \ldots, W_k) \leq x\} \geq \gamma\} \ .$$

Fix $0 \leq \beta \leq \alpha$. The test of (S.1) that rejects H_0 if $T > b(1-\alpha+\beta, \tilde{\mu})$ satisfies (S.3).

Remark S.1.2. Although we are unable to establish that the left-hand side of (S.3) equals α, we are able to establish that the left-hand side of (S.3) is at least $\alpha - \beta$. To see this, simply note that $b(1-\alpha+\beta, \tilde{\mu}) \leq b(1-\alpha+\beta, 0)$, so

$$\sup_{\mu \in \Omega_0} P_\mu\{T > b(1-\alpha+\beta, \tilde{\mu})\} \geq P_0\{T > b(1-\alpha+\beta, 0)\} = \alpha - \beta \ . \blacksquare$$

Remark S.1.3. As emphasized above, an attractive feature of our procedure is that the "largest" value of μ in $\Omega_0 \cap M(1-\beta)$ may be determined explicitly. This follows from our particular choice of initial confidence region for μ, namely, from its rectangular shape. If, for example, we had instead chosen $M(1-\beta)$ to be the usual confidence ellipsoid, then there might not even be a "largest" value of μ in $\Omega_0 \cap M(1-\beta)$, and one would have to compute

$$\sup_{\mu \in \Omega_0 \cap M(1-\beta)} b(1-\alpha+\beta, \mu) \ .$$

This problem persists even if the initial confidence region is chosen by inverting tests based on the likelihood ratio statistic (S.5) despite the resulting confidence region being monotone decreasing in the sense that if x lies in the region, then so does y whenever $y_j \leq x_j$ for all $1 \leq j \leq k$. $\blacksquare$

Remark S.1.4. In some cases, it may be desired to test the null hypothesis that $\mu \in \tilde{\Omega}_0$, where

$$\tilde{\Omega}_0 = \{\mu : \mu_j = 0 \text{ for } j \in J_1, \mu_j \leq 0 \text{ for } j \in J_2\}$$

and J_1 and J_2 form a partition of $\{1, \ldots, k\}$. Such a situation may be accommodated in the framework described above simply by writing $\mu_j = 0$ as $\mu_j \leq 0$ and $-\mu_j \leq 0$, but the resulting procedure may be improved upon by exploiting the additional structure of the null hypothesis. In particular, Theorem S.1.2 remains valid if T is only required to be nondecreasing in its $|J_2|$ arguments with $j \in J_2$ and $\tilde{\mu}$ is replaced by the vector whose jth component is equal to 0 for $j \in J_1$ and $\min\{W_j + \tilde{K}^{-1}(1-\beta), 0\}$ for $j \in J_2$, where $\tilde{K}^{-1}(1-\beta)$ is the $1-\beta$ quantile of the distribution

$$\tilde{K}(x) = P_\mu\left\{\max_{j \in J_2}(\mu_j - W_j) \leq x\right\} . \blacksquare$$

Remark S.1.5. In the context of the Gaussian model considered in this section, it is instructive for comparison purposes to consider a parametric counterpart to the nonparametric method of Andrews and Barwick (2012a). To describe their approach, fix $\kappa < 0$. Let $\hat{\mu}$ be the k-dimensional vector whose jth component equals zero if $W_j > \kappa$ and $-\infty$ otherwise (or, for practical purposes, some very large negative number). Define the "size correction factor"

$$\hat{\eta} = \inf\left\{\eta > 0 : \sup_{\mu \in \Omega_0} P_\mu\{T > b(1-\alpha, \hat{\mu}) + \eta\} \leq \alpha\right\} . \tag{S.8}$$

The proposed test of (S.1) then rejects H_0 if $T > b(1-\alpha, \hat{\mu}) + \hat{\eta}$. The addition of $\hat{\eta}$ is required because, in order to allow the asymptotic framework to better reflect the finite-sample situation, the authors do not allow κ to tend to zero with the sample size n. Note that the computation of $\hat{\eta}$ as defined in (S.8) is complicated by the fact that there is no explicit solution to the supremum in (S.8). One must, therefore, resort to approximating the supremum in (S.8) in some fashion. Andrews and Barwick (2012a) propose to approximate $\sup_{\mu \in \Omega_0} P_\mu\{T > b(1-\alpha, \hat{\mu}) + \eta\}$ with $\sup_{\mu \in \tilde{\Omega}_0} P_\mu\{T > b(1-\alpha, \hat{\mu}) + \eta\}$, where $\tilde{\Omega}_0 = \{-\infty, 0\}^k$. Andrews and Barwick (2012a) provide an extensive simulation study, but no proof, in favor of this approximation. Even so, the problem remains computationally demanding and, as a result, the authors only consider situations in which $k \leq 10$ and $\alpha = .05$. In contrast, our two-step procedure is simple to implement even when k is large, as it does not require optimization over Ω_0, and has proven size control for any value of α (thereby allowing, among other things, one to compute a p-value as the smallest value of α for which the null hypothesis is rejected). In the nonparametric setting considered below, where the underlying covariance matrix is also unknown, further approximations are required to implement the method of Andrews and Barwick (2012a). See Remark 2.6 for related discussion. $\blacksquare$

Remark S.1.6. Let $\phi_{\alpha,\beta}$ be the test as described in Theorem S.1.2. Similar to the approach of Andrews and Barwick (2012a), one can determine β to maximize (weighted) average power. In the parametric context considered in this section, one can achieve this exactly modulo simulation error. To describe how, let $\mu_1, \ldots, \mu_d$ be alternative values in Ω_1, and let $w_1, \ldots, w_d$ be nonnegative weights that add up to one. Then, β can be chosen to maximize

$$\sum_{i=1}^{d} w_i \mathbb{E}_{\mu_i}[\phi_{\alpha,\beta}] .$$

This can be accomplished by standard simulation from $N(\mu_i, \Sigma)$ and discretizing β between 0 and α. The drawback here is the specification of the μ_i and w_i. In our simulations, we have found that a reasonable choice is simply $\beta = \alpha/10$. $\blacksquare$

A Appendix

Proof of Theorem S.1.1. For $1 \leq j \leq k$, let e_j be the jth unit basis vector having a 1 in the jth coordinate. To determine $\bar{\mu}$ for the given a, we must minimize

$$f(\mu) = (\mu - a)'\Sigma^{-1}(\mu - a)$$

over $\mu \in \Omega_0$. Note that

$$\frac{\partial f(\mu)}{\partial \mu_j} = 2(\mu - a)'\Sigma^{-1}e_j \ .$$

First of all, we claim that the minimizing $\bar{\mu}$ cannot have all of its components negative. This follows because, if it did, the line joining the claimed solution and a itself would intersect the boundary of Ω_0 at a point with a smaller value of $f(\mu)$. Therefore, the solution $\bar{\mu}$ must have at least one zero entry.

Suppose that $\bar{\mu}$ is the solution and that $\bar{\mu}_j = 0$ for $j \in J$, where J is some nonempty subset of $\{1, \ldots, k\}$. Let $f_J(\mu) = f(\mu)$ viewed as a function of μ_j with $j \notin J$ and with $\mu_j = 0$ for $j \in J$. Then, the solution to the components $\bar{\mu}_j$ with $j \notin J$ (if there are any) must be obtained by setting partial derivatives equal to zero, leading to the solution of the equations

$$(\mu - a)'\Sigma^{-1}e_j = 0 \ \forall j \notin J$$

with μ_j fixed at 0 for $j \in J$. Now, the MP test for testing $\bar{u}$ against a rejects for large values of $W'\Sigma^{-1}(a - \bar{u})$, which is a linear combination of $W_1, \ldots, W_k$. The coefficient multiplying W_j is $e_j'\Sigma^{-1}(a - \bar{u})$. But for $j \notin J$, this coefficient is zero by the gradient calculation above.

Next we claim that for $j = 1, \ldots, k$, the coefficient of W_j is nonnegative. Fix j. Consider $f(\mu)$ as a function of μ_j alone with the other components fixed at the claimed solution for $\bar{\mu}$. If the derivative with respect to μ_j at 0 were positive, i.e.,

$$(\bar{\mu} - a)'\Sigma^{-1}e_j > 0 \ ,$$

then the value of μ_j could decrease and result in a smaller minimizing value for $f(\mu)$. Therefore, it must be the case that

$$(a - \bar{\mu})'\Sigma^{-1}e_j \geq 0 \ ;$$

the left-hand side is precisely the coefficient of W_j.

Thus, the solution $\bar{\mu}$ has the property that, for testing $\bar{\mu}$ against a, the MP test rejects for large $\sum_{1 \leq j \leq k} c_j W_j$ such that $\bar{\mu}_j = 0$ implies $c_j \geq 0$ and $\bar{\mu}_j < 0$ implies $c_j = 0$. This property allows us to prove that $\bar{\mu}$ is least favorable. Indeed, if the critical value c is determined so that the test is level α under $\bar{\mu}$, then for $\mu \in \Omega_0$,

$$P_\mu\Big\{\sum_{j \in J} c_j W_j > c\Big\} \leq P_0\Big\{\sum_{j \in J} c_j W_j > c\Big\} = P_{\bar{\mu}}\Big\{\sum_{j \in J} c_j W_j > c\Big\}.$$

The first inequality follows by monotonicity and the second one by the fact that $\bar{\mu}_j = 0$ for $j \in J$. The least favorable property now follows byLehmann and Romano (2005, Theorem 3.8.1).

The remainder of the proof is obvious. ■

Proof of Theorem S.1.2 First note that $b(\gamma, \mu)$ is nondecreasing in μ, since T is nondecreasing in its arguments. Fix any μ with $\mu_i \leq 0$. Let E be the event that $\mu \in M(1-\beta)$. Then, the Type I error satisfies

$$P_\mu\{\text{reject } H_0\} \leq P_\mu\big\{E^c\big\} + P_\mu\big\{E \cap \{\text{reject } H_0\}\big\} = \beta + P_\mu\big\{E \cap \{\text{reject } H_0\}\big\} \ .$$

But when the event E occurs and H_0 is rejected — so that $T > b(1-\alpha+\beta, \tilde{\mu})$ — then the event $T > b(1-\alpha+\beta, \mu)$ must occur, since $b(1-\alpha+\beta, \mu)$ is nondecreasing in μ and $\mu \leq \tilde{\mu}$ when E occurs. Hence, the Type I error is bounded above by

$$\beta + P_\mu\big\{T > b(1-\alpha+\beta, \mu)\big\} \leq \beta + \big(1 - (1-\alpha+\beta)\big) = \alpha \ . \ ■$$

References

Andrews, D. (2012). Similar-on-the-boundary tests for moment inequalities exist, but have poor power. Technical report, Department of Economics, Yale University.

Andrews, D. W. K. and Barwick, P. J. (2012a). Inference for parameters defined by moment inequalities: A recommended moment selection procedure. *Econometrica*, 80(6):2805–2826.

Andrews, D. W. K. and Barwick, P. J. (2012b). Inference for parameters defined by moment inequalities: A recommended moment selection procedure. *Econometrica Supplemental Material*, 80(6). Available at `http://www.econometricsociety.org/`.

Lehmann, E. L. (1952). Testing multiparameter hypotheses. *Annals of Mathematical Statistics*, 23:541–552.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer, New York, third edition.