Approximately Most Powerful Tests for Moment Inequalities

Richard C. Chiburis*
Department of Economics, Princeton University

September 26, 2008

Abstract

The existing literature on testing moment inequalities has focused on finding appropriate critical values for tests based on a given objective function, but the objective functions themselves are chosen in an ad hoc manner, and the power of a test depends on the objective function that was used. In contrast, we apply a general algorithm to approximate tests that maximize weighted average power against the alternative hypothesis. The general algorithm is computationally feasible only for low-dimensional problems, but we present another test that is feasible in higher dimensions. This test exactly satisfies an intuitive optimality condition, and simulations show that it outperforms other tests in the literature and comes close to the maximum weighted average power.

JEL classifications: C12, C30.

1 Introduction

Many econometric models are only partially identified, meaning that not just a single value of the parameter of interest θ_0, but rather many values, are consistent with the distribution underlying the data. Examples of partially identified models include treatment effects models and other models involving endogenous variables, missing data, or multiple equilibria. Often partially identified models can be written as a set of moment inequalities

$$\mathrm{E}\left[m\left(Y,\theta_0\right)\right] \leq 0, \tag{1}$$

where Y is the random variable from which the data is sampled, and m is a known vector-valued function. We are interested in testing whether a particular value of θ_0 satisfies (1).

*I would like to thank Andrés Aradillas-López, Bo Honoré, Ulrich Müller, Sam Schulhofer-Wohl, Neşe Yıldız, and participants at the Copenhagen Microeconometrics Summer Workshop for helpful comments.

While the literature on inference for moment-inequality models is growing rapidly, little attention has been paid thus far to finding most powerful tests, which we define as tests that maximize weighted average power against a given distribution over the alternative hypothesis; typically no test is uniformly most powerful against all alternatives. Instead, the available tests are based on ad-hoc objective functions.

Rosen (2007) defines a test for moment inequalities based on computing the worst-case critical value within the null hypothesis (1) of a particular objective function. Typically the worst-case distribution of the objective function within the null occurs at the point where the maximal number of inequalities (1) are satisfied with equality. The usage of this worst-case critical value often makes Rosen's test unnecessarily conservative away from the worst-case point.

Andrews and Soares (2007) improve the power of Rosen's test using a technique called generalized moment selection (GMS), which selects particular inequalities that are very likely satisfied by a wide margin and recomputes the worst-case critical value *without* those inequalities. This works because non-binding inequalities are unlikely to contribute to the value of a certain class of objective functions. However, Andrews and Soares do not consider the optimal choice of objective function. Furthermore, Andrews and Soares show that their test has *uniformly* asymptotically correct size, whereas Rosen's test is only pointwise asymptotically valid. As shown by Imbens and Manski (2004), tests that are not uniformly justified are often undersized in finite samples. While Andrews and Soares' test is also undersized in finite samples because with small probability some inequalities may be dropped from the objective function incorrectly, the magnitude of this error appears to be negligible.

Chernozhukov, Hong, and Tamer (2007) develop moment inequality tests based on critical values computed via subsampling. For a given objective function, subsampling can improve upon the worst-case critical value because it captures the behavior of the objective function locally to the actual data. However, Andrews and Soares provide local asymptotic results showing that GMS is more powerful than subsampling, even asymptotically. Again, the tests that are compared are based on a particular ad-hoc objective function.

We also consider *conditional* moment inequality models, in which the data can be separated into two random variables X and Y such that

$$\mu(x) \ \leq 0 \text{ for all } x \in \mathcal{X}, \tag{2}$$

where $\mu(x) = \mathrm{E}\left[m(X, Y, \theta_0) \mid X = x\right]$.

Semiparametric models that are partially identified often can be written in terms of conditional moment inequalities. Manski and Tamer (2002) study a regression model with an interval observed outcome:

$$\mathrm{E}\left[Y_i \mid X_i = x\right] = x'\theta_0.$$

The econometrician observes X_i and bounds Y_{Li}, Y_{Ui} such that $Y_{Li} \leq Y_i \leq Y_{Ui}$. Then we have two conditional moment inequalities:

$$\begin{aligned}\mathrm{E}\left[Y_{Li}-X_i'\theta_0 \mid X_i=x\right] &\leq 0 \text{ for all } x\in\mathcal{X}\\ \mathrm{E}\left[Y_{Ui}-X_i'\theta_0 \mid X_i=x\right] &\geq 0 \text{ for all } x\in\mathcal{X}.\end{aligned}$$

Chiburis (2008) shows that a semiparametric model for bounds on treatment effects reduces to a system of conditional moment inequalities.

A time-series example is submartingale testing. Testing whether a stochastic process is a submartingale can also be viewed as a single conditional moment inequality:

$$\mathrm{E}\left[Y_i-X_i \mid X_i=x\right]\leq 0 \text{ for all } x\in\mathcal{X}$$

where $Y_i = X_{i+1.}$

The usual first step in approaching conditional moment inequalities is to observe that (2) implies infinitely many unconditional moment inequalities:

$$\mathrm{E}\left[m(X,Y,\theta_0)^\top h(X)\right]\leq 0 \tag{3}$$

for any *nonnegative* function $h(x)$. Existing tests for conditional equalities or inequalities are all based on (3) for a particular set of functions h. However, if only a finite, fixed set of functions h is used, the resulting test will be inconsistent against some alternatives.

For the case of conditional moment *equalities*, Bierens (1990) proposes a test based on an infinite class of functions h, such as exponential functions (similar to a Fourier transform) or polynomials of various degrees. Also for equalities, Domínguez and Lobato (2004) present an estimator using indicator functions h for intervals over the empirical dataset. Conditional moment inequalities are more difficult to test than conditional moment equalities because h must be nonnegative for inequalities, which precludes the use of an orthogonal class of h functions, and also because there are many different functions μ that satisfy (2).

Both Khan and Tamer (2006) and Kim (2008) adapt Domínguez and Lobato's approach to inequalities, although Khan and Tamer focus on a particular application and the special case in which their model is point-identified. Kim's test uses an objective function based on sums of modified-χ^2 test statistics over all possible intervals of the empirical dataset. The intuition for this is that he wants to have power against localized violations $\mu(x)>0$ for each possible range of x. Define $m_i = m(X_i,Y_i,\theta)$. The population version of Kim's objective function for scalar m and X is

$$Q(\theta)=\mathrm{E}_{X_j,X_k}\left[\frac{\mathrm{E}_{X_i}\left[m_i\mathbf{1}\left\{X_j\leq X_i\leq X_k\right\}\right]_+^2}{Var[m_i\mathbf{1}\left\{X_j\leq X_i\leq X_k\right\}]}\right],$$

and the sample objective function is

$$\hat{Q}_n(\theta) = \tfrac{1}{n(n-1)} \sum_{j \neq k} \left[\frac{\left(\frac{1}{n}\sum_{i=1}^{n} m_i \mathbf{1}\left\{x_j \leq x_i \leq x_k\right\}\right)_+^2}{\frac{1}{n}\sum_{i=1}^{n} \left(m_i \mathbf{1}\left\{x_j \leq x_i \leq x_k\right\}\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n} m_i \mathbf{1}\left\{x_j \leq x_i \leq x_k\right\}\right)^2} \right]. \tag{4}$$

This sum emphasizes local deviations by dividing each term by its variance.

The critical value of $\hat{Q}_n$ under the null (2) actually depends on the particular function $\mu(x) = \mathrm{E}\left[m(X, Y, \theta_0) \mid X = x\right]$ within the null hypothesis. The worst case is usually $\mu(x) = 0$ for all $x \in \mathcal{X}$. For standard normal disturbances, the 5% worst-case critical value of $n\hat{Q}_n$ is approximately 1.7 for n sufficiently large (≥ 100). To do better than using the worst-case critical value, Kim adapts the subsampling approach of Chernozhukov, Hong, and Tamer (2007), but no results are presented. Kim also considers estimation of the model, defining a set estimator of the identified set and computing its convergence rate in the Hausdorff metric.

Khan and Tamer (2006) and Kim (2008) do not show why their particular choice of objective function is the best one. For example, it seems arbitrary to check for deviations over all intervals of the data but not over the set complements of these intervals. The relative weightings of the intervals are also ad hoc.

In this paper, we present in Section 2 an algorithm for approximating *most powerful* tests for moment inequalities against a particular distribution over the alternative hypothesis. The algorithm is very general but only computationally feasible for low-dimensional data. To remedy this problem, in Section 3 we develop a test for moment inequalities that can be computed for high-dimensional data. In a subset of cases we demonstrate uniform convergence in size and show that the test satisfies an intuitive optimality condition. In Section 4 we show how the test can be applied to conditional moment inequalities, and we compare finite-sample size and power of our test with other available tests. Section 5 concludes.

2 Approximately most powerful tests

We observe n i.i.d. draws $\{(Y_i)\}_{i=1}^{n}$ of the random variable $Y \in \mathcal{Y}$. Our goal is to construct a powerful test of the moment inequality hypothesis $\mathrm{E}\left[m\left(Y, \theta\right)\right] \leq 0$ for a given value of θ, where $m : \mathcal{Y} \times \Theta \rightarrow \mathbb{R}^K$ is a known function. Defining the $K \times 1$ vector $\mu = \mathrm{E}\left[m\left(Y, \theta\right)\right]$, we can rewrite the null hypothesis as

$$H_0 : \mu_k \leq 0 \text{ for all } k \in \{1, \ldots, K\}. \tag{5}$$

To start, we set the alternative hypothesis to be the set complement of H_0:

$$H_a : \text{ There exists } k \in \{1, \ldots, K\} \text{ such that } \mu_k > 0.$$

Moment inequalities make for a difficult testing problem because both the null and alternative hypotheses are composite; there are many vectors μ which satisfy the null hypothesis with varying degrees of tightness in various coordinates, and there are also many ways in which the null hypothesis can be violated. But even when the null and alternative hypotheses are composite and there is no uniformly most powerful test, a most powerful test is still well-defined for a given *weighting* $\pi(\mu)$ over H_a (Andrews and Ploberger 1994). We then consider the most powerful test of the null (5) against the simple alternative

$$H_\pi : \mu \sim \pi. \tag{6}$$

We want to maximize power over the alternative such that proper size over all points in the null is maintained. There is no known way to solve this analytically, but we can approximate the solution numerically in some simple cases.

2.1 Algorithm for approximating a most powerful test

We present a general algorithm can be used to approximate the most powerful test if K is small. While other general algorithms for most powerful tests exist based on finding the least favorable distribution over the null, e.g. Müller and Watson (2008), Algorithm 2.1 has the advantage of being a linear program rather than a nonlinear optimization. We discuss a duality connection to these other algorithms in Appendix A.1, and in Appendix A.2 we develop a faster, linear-program version of Müller and Watson's algorithm.

We add the following assumption:

Assumption 1 *Assume that $m(Y,\theta)$ has finite variance.*

Define $\Sigma = \text{Var}\left[m(Y,\theta)\right]$. An unbiased, consistent estimator of Σ is

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(m(Y_i,\theta)-\bar{m})(m(Y_i,\theta)-\bar{m})^\top,$$

where the $K \times 1$ random vector $\bar{m}$ is defined as

$$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m(Y_i,\theta).$$

Since a central limit theorem yields

$$\sqrt{n}\left(\bar{m}-\mu\right) \Rightarrow \mathcal{N}(0,\Sigma), \tag{7}$$

we will use the asymptotic approximation

$$\bar{m} \sim \mathcal{N}(\mu,\bar{\Sigma}) \tag{8}$$

for the distribution of $\bar{m}$, where $\bar{\Sigma} = \frac{1}{n}\hat{\Sigma}$. We now present a general algorithm that approximates the most powerful test of H_0 (5) based on the value of $\bar{m}$:[1]

Algorithm 2.1 *Algorithm for approximating the most powerful size-α test of the composite null hypothesis H_0 (5) against the simple alternative H_π (6) under the distributional assumption (8).*

1. *Partition $\mathbb{R}^K$ into a grid composed of a finite set S of regions s. This may be done by defining a finite rectangular grid covering the area of interest near the origin and adding an additional region that encompasses all of the space outside of the grid.*

2. *Define a finite grid of points $\{\mu^j\}_{j=1}^J$ over H_0 for the purpose of checking size at those points.*[2]

3. *Compute the probability π_s under H_π that $\bar{m}$ would have fallen within each grid region s. Use H_π for the distribution of μ and approximate $\bar{m} \sim \mathcal{N}(\mu, \bar{\Sigma})$.*

4. *Compute the probability $\nu_{j,s}$ under each null hypothesis point μ^j that m_i would have fallen within each grid region s. Use the approximation $\bar{m} \sim \mathcal{N}(\mu^j, \bar{\Sigma})$.*

5. *Solve the linear program*

$$\begin{aligned} \max_{\{\phi_s\}} \quad & \sum_{s \in S} \pi_s \phi_s \\ s.t. \quad & \sum_{s \in S} \nu_{j,s} \phi_s \leq \alpha \textit{ for all } j \in \{1, \ldots, J\} \\ & 0 \leq \phi_s \leq 1 \textit{ for all } s \in S. \end{aligned} \tag{9}$$

6. *Let $\hat{s}$ be the region that contains the actual $\bar{m}$ observed in the data. Reject H_0 with probability $\phi_{\hat{s}}$.*

In words, the linear program assigns a rejection probability ϕ_s to each square s to maximize the power against H_π such that the size at each μ^j in the null is at most α. Since this is a linear programming problem, the maximum can be found easily, and the only source of approximation is the discrete grid. The idea of Algorithm 2.1 is quite general, and it can easily be adapted to many applications beyond moment inequalities.

[1] The optimality of this test is based on the asymptotic normal approximation (7) and the assumption that $\hat{\Sigma} = \Sigma$, in which case the likelihood of μ is a function only of $\bar{m}$. If the exact distribution of $m(Y_i, \theta) - \mu$ is known, then the algorithm can be improved by using the exact distribution to compute likelihoods.

[2] In fact, simulations show that it suffices to check size only on the boundary of H_0, as long as H_π has no overlap with H_0.

2.2 Approximate optimality

Due to the discretization and the normal approximation, Algorithm 2.1 only *approximates* the most powerful test of H_0 against H_π. The three sources of error in the approximation are (1) that the size is only controlled at a finite set of points, (2) that there could be a more powerful test if the rejection probability were not held constant within each region s, and (3) the asymptotic normal approximation for the distribution of $\bar{m} - \mu$.

2.2.1 Size control

Algorithm 2.1 controls size explicitly only at certain grid points. What about the rest of the points in H_0 within the area covered by the grid?[3] First consider any two grid points $\mu, \mu' \in H_0$ and a point between them $\tilde{\mu} = \lambda\mu + (1-\lambda)\mu'$ for some $\lambda \in (0,1)$. If the size at μ and μ' is α, the size for a simple null ξ with weight λ on μ and $1-\lambda$ on μ' is also α, since $\mathrm{E}_\xi\left[\phi(\bar{m})\right] = \lambda\,\mathrm{E}_\mu\left[\phi(\bar{m})\right] + (1-\lambda)\,\mathrm{E}_{\mu'}\left[\phi(\bar{m})\right]$. We want to find a bound on $\mathrm{E}_{\tilde{\mu}}\left[\phi(\bar{m})\right]$. We use the fact that the difference between the two sizes cannot exceed the *total variation distance* between the distributions of $\bar{m}$ in the two scenarios. The total variation distance between two distributions F_1 and F_2 is the maximum difference in the probabilities that the two distributions can assign to the same event, and it is equal to $\int \left[F_1(z) - F_2(z)\right]_+ dz$, where $[x]_+ = \max\{x, 0\}$. The following lemma bounds this total variation distance:

Lemma 2.1 *The total variation distance between $\mathcal{N}(\tilde{\mu}, \bar{\Sigma})$ and the mixture of normals $\lambda\mathcal{N}(\mu, \bar{\Sigma}) + (1-\lambda)\mathcal{N}(\mu', \bar{\Sigma})$ is bounded above by $\frac{\Phi'(1)}{4\omega}\left\|\mu - \mu'\right\|^2$, where $\Phi(\cdot)$ is the standard univariate Gaussian distribution function, ω is the smallest eigenvalue of $\bar{\Sigma}$, and $\|\cdot\|$ is the Euclidean norm.*

Proof. Let $d = \|\mu - \mu'\|$. Since μ, μ', and $\tilde{\mu}$ lie on a line, it is sufficient to consider variation along the dimension of that line. The marginal distribution of $(\bar{m} - \mu)$ along any line is univariate normal with variance at least ω. Without loss of generality we may use $\mu = -(1-\lambda)d$ and $\mu' = \lambda d$, so that $\tilde{\mu} = 0$, and we assume the worst-case variance ω. Then the total variation distance Δ between $\mathcal{N}(0, \omega)$ and the mixture $\lambda\mathcal{N}(-(1-\lambda)d, \omega) + (1-\lambda)\mathcal{N}(\lambda d, \omega)$ is

$$\begin{aligned} \Delta &= \frac{1}{\sqrt{2\pi\omega}} \int_{-\infty}^{\infty} \left[e^{-\frac{1}{2\omega}z^2} - \lambda e^{-\frac{1}{2\omega}(z+(1-\lambda)d)^2} - (1-\lambda)e^{-\frac{1}{2\omega}(z-\lambda d)^2}\right]_+ dz \\ &= \frac{1}{\sqrt{2\pi\omega}} \int_{a_{\lambda,d}}^{b_{\lambda,d}} \left(e^{-\frac{1}{2\omega}z^2} - \lambda e^{-\frac{1}{2\omega}(z+(1-\lambda)d)^2} - (1-\lambda)e^{-\frac{1}{2\omega}(z-\lambda d)^2}\right) dz \end{aligned}$$

[3] Outside of the grid, there is no guarantee of size control, but this problem can be remedied by forcing the test to accept in the outer region described in step 1 of Algorithm 2.1. This ensures that the test is conservative away from the region covered by the grid.

for some $a_{\lambda,d}, b_{\lambda,d}$. The derivative with respect to d is[4]

$$\begin{aligned}
& \frac{\lambda(1-\lambda)}{\sqrt{2\pi\omega}} \int_{a_{\lambda,d}}^{b_{\lambda,d}} \left(\frac{z+(1-\lambda)d}{\omega} e^{-\frac{1}{2\omega}(z+(1-\lambda)d)^2} - \frac{z-\lambda d}{\omega} e^{-\frac{1}{2\omega}(z-\lambda d)^2} \right) dz \\
= & \frac{\lambda(1-\lambda)}{\sqrt{2\pi\omega}} \left(e^{-\frac{1}{2\omega}(z-\lambda d)^2} - e^{-\frac{1}{2\omega}(z+(1-\lambda)d)^2} \right) \Big|_{a_{\lambda,d}}^{b_{\lambda,d}} \\
= & \frac{\lambda(1-\lambda)}{\sqrt{\omega}} \left[\left(\Phi'\left(\tfrac{b_{\lambda,d}-\lambda d}{\sqrt{\omega}}\right) - \Phi'\left(\tfrac{b_{\lambda,d}+(1-\lambda)d}{\sqrt{\omega}}\right) \right) - \left(\Phi'\left(\tfrac{a_{\lambda,d}-\lambda d}{\sqrt{\omega}}\right) - \Phi'\left(\tfrac{a_{\lambda,d}+(1-\lambda)d}{\sqrt{\omega}}\right) \right) \right] \\
\leq & \frac{2\lambda(1-\lambda)\left|\Phi''(1)\right| d}{\omega}, \text{ since } \left|\Phi''(z)\right| \text{ is maximized at } z = \pm 1 \\
\leq & \frac{\Phi'(1) d}{2\omega}, \text{ since } \lambda(1-\lambda) \leq \frac{1}{4} \text{ and } \left|\Phi''(1)\right| = \Phi'(1).
\end{aligned}$$

Integrating this from 0 to d, we get $\Delta \leq \frac{\Phi'(1)d^2}{4\omega}$. ■

If $K > 1$, then not all points in H_0 are directly between grid points at which Algorithm 2.1 controls size. However, we can still bound size at all points in H_0 by applying Lemma 2.1 repeatedly, as shown in the following lemma:

Lemma 2.2 *Suppose that a rectangular grid with distance d between adjacent grid points is set up over H_0, and a test ϕ has size at most α at all grid points. Then ϕ has size at most*

$$\alpha + \frac{K\Phi'(1)d^2}{4\omega} \tag{10}$$

over all of H_0.

Proof. Let μ be any point in H_0. We prove the lemma by induction on K.

If $K = 1$, then μ must lie between two grid points, at which the size is α, so by Lemma 2.1, the size at μ is at most $\alpha + \frac{\Phi'(1)d^2}{4\omega}$.

If $K > 1$, define $\underline{\mu}$ such that $\underline{\mu}_i = \mu_i$ for all $i \in \{1, \ldots, K-1\}$, and $\underline{\mu}_K$ is the largest grid coordinate that satisfies $\underline{\mu}_K \leq \mu_K$. Similarly, define $\overline{\mu}$ such that $\overline{\mu}_i = \mu_i$ for all $i \in \{1, \ldots, K-1\}$, and $\overline{\mu}_K$ is the smallest grid coordinate that satisfies $\overline{\mu}_K > \mu_K$. By induction, we can apply the lemma along the first $K-1$ dimensions to obtain that the power at each of $\underline{\mu}$ and $\overline{\mu}$ is $\alpha + \frac{(K-1)\Phi'(1)d^2}{4\omega}$. Since μ lies between $\underline{\mu}$ and $\overline{\mu}$, which are distance d apart, Lemma 2.1 tells us that the size at μ is at most $\alpha + \frac{K\Phi'(1)d^2}{4\omega}$. ■

In most cases, the K factor in Lemma 2.2 can be replaced by $(K-1)$, since often it is only necessary to check size on the boundary of H_0, and the boundary is $(K-1)$-dimensional. Lemma 2.2 assumes that we can set up a grid over all of H_0, but in fact it is only possible to set up a grid over a finite region of H_0. However, it is usually possible to deduce the shape of the optimal test over all of H_0 by examining the

[4] The terms related to the differentiation of the limits of integration $a_{\lambda,d}$ and $b_{\lambda,d}$ disappear since the integrand is zero at those points.

optimal test over a finite grid near the origin, since the interesting behavior typically occurs near the origin.

Note that these results are uninformative when $\bar{\Sigma}$ is singular, so that the minimum eigenvalue ω is zero. In such cases, one can still transform the problem by defining the grid in the step 1 of Algorithm 2.1 over the domain Y of $\mathcal{Y}$ rather than the domain $\mathbb{R}^K$ of $\bar{m}$. Often the covariance matrix of Y will be nonsingular even when the covariance matrix of $\bar{m}$ is singular.

While it may seem problematic in (10) that ω moves in proportion with $1/n$, the variance of $\bar{m}$ shrinks as well. Hence the boundaries of the grids may be shrunk in proportion with $1/\sqrt{n}$ with no loss in precision, and d may be set in proportion to $1/\sqrt{n}$ without losing computational speed. The net result is that the computation time necessary to achieve a given precision in the size does not depend on n.

2.2.2 Power approximation

Let g_π be the probability density of $\bar{m}$ under π. Algorithm 2.1 computes the power of tests against a discretized version $\tilde{g}_{\pi,d}$ of g_π. How does the power of a test against $\tilde{g}_{\pi,d}$ compare to the power of a test against g_π? Once again, the difference in the powers is at most the total variation distance between the two distributions.

Let ϕ be the test produced by Algorithm 2.1, and let $\tilde{\phi}$ be any size-α test of H_0. By the construction of ϕ, no size-α test has better power against $\tilde{g}_{\pi,d}$ than ϕ, so

$$\int_{\mathbb{R}^K} \phi(z)\tilde{g}_{\pi,d}(z)dz \geq \int_{\mathbb{R}^K} \tilde{\phi}(z)\tilde{g}_{\pi,d}(z)dz. \tag{11}$$

Furthermore, since ϕ is constant within each grid cube, and $\tilde{g}_{\pi,d}(z)$ is defined so that the integral of $\tilde{g}_{\pi,d}(z)$ within each cube equals the integral of $g_\pi(z)$ within that cube,

$$\int_{\mathbb{R}^K} \phi(z)\tilde{g}_{\pi,d}(z)dz = \int_{\mathbb{R}^K} \phi(z)g_\pi(z)dz. \tag{12}$$

Combining (11) and (12),

$$\int_{\mathbb{R}^K} \phi(z)g_\pi(z)dz \geq \int_{\mathbb{R}^K} \tilde{\phi}(z)\tilde{g}_{\pi,d}(z)dz.$$

By the definition of total variation distance,

$$\int_{\mathbb{R}^K} \phi(z)g_\pi(z)dz \geq -\Delta_{\tilde{g}_{\pi,d},g_\pi} + \int_{\mathbb{R}^K} \tilde{\phi}(z)g_\pi(z)dz,$$

where $\Delta_{\tilde{g}_{\pi,d},g_\pi}$ is the total variation distance between g_π and $\tilde{g}_{\pi,d}$.

All that remains is to approximate $\Delta_{\tilde{g}_{\pi,d},g_\pi}$. Suppose that the grid $\mathbb{R}^K$ is composed of cubical regions with side length d.[5] If g_π is smooth, then for sufficiently small d,

[5] In practice we have suggested creating a finite number of grid regions by partitioning a bounded area of $\mathbb{R}^K$ into a grid, and then add one more region that includes everything outside the grid. The total variation distance given here will be approximately valid if the probability under g_π on the outer region is small.

the total variation distance between g_π and a flat (uniform) distribution within a cube is well approximated by the area of a triangle with height $\frac{d}{2}\left\|g_\pi'(z)\right\|$ and base $\frac{d^K}{2}$, which is $\frac{d^{K+1}}{8}\left\|g_\pi'(z)\right\|$. Integrating this over the entire grid, we get the following approximation for the total variation distance $\Delta_{\tilde{g}_{\pi,d},g_\pi}$ between $\tilde{g}_{\pi,d}$ and g_π:

$$\Delta_{\tilde{g}_{\pi,d},g_\pi} \approx \frac{d}{8}\int_{\mathbb{R}^K}\left\|g_\pi'(z)\right\| dz.$$

For example, if g_π is standard univariate normal, $\Delta_{\tilde{g}_{\pi,d},g_\pi} \approx \frac{d}{4\sqrt{2\pi}}$, and if g_π is standard bivariate normal, $\Delta_{\tilde{g}_{\pi,d},g_\pi} \approx \frac{d\sqrt{2\pi}}{16}$. Note that this is an approximation rather than an upper bound, but numerical calculations for the standard bivariate normal distribution show that the approximation is quite good.

Therefore, the power of ϕ against g_π is within $\Delta_{\tilde{g}_{\pi,d},g_\pi} \approx \frac{d}{8}\int_{\mathbb{R}^K}\left\|g_\pi'(z)\right\| dz$ of the power of the most powerful size-α test against g_π.

2.2.3 Normal approximation

All of our analysis so far has been based on the asymptotic approximation $\bar{m} \sim \mathcal{N}(\mu, \bar{\Sigma})$. Let F_n be the true distribution of $\bar{m} - \mu$.

Lemma 2.3 *The total variation distance between $\mathcal{N}(\mu, \bar{\Sigma})$ and F_n approaches zero as $n \to \infty$.*

Proof. By (7), the total variation distance between $\sqrt{n}F_n$ and $\mathcal{N}(0, \Sigma)$ approaches zero as $n \to \infty$. Since $\hat{\Sigma} \xrightarrow{p} \Sigma$, the total variation distance between $\mathcal{N}(0, \Sigma)$ and $\mathcal{N}(0, \hat{\Sigma})$ approaches zero as well. Then by the triangle inequality, the total variation distance between $\sqrt{n}F_n$ and $\mathcal{N}(0, \hat{\Sigma})$ approaches zero. This is equal to the total variation distance between F_n and $\mathcal{N}(0, \bar{\Sigma})$, since total variation distance is invariant to multiplication by a constant. ■

It follows from Lemma 2.3 that any error due to the asymptotic normal approximation in the size or power of the test produced by Algorithm 2.1 approaches zero as $n \to \infty$.

2.2.4 Summary

We now combine the results of Sections 2.2.1, 2.2.2, and 2.2.3. Let ϕ be the test produced by Algorithm 2.1. We know that ϕ has size at most $\alpha + \frac{\Phi'(1)Kd^2}{4\omega}$, and power within approximately $\frac{d}{8}\int_{\mathbb{R}^K}\left\|g_\pi'(z)\right\| dz$ of the most powerful size-α test against g_π. Therefore, by making making the grid width d smaller, one can use Algorithm 2.1 to get arbitrarily close to an optimal test, except for error due to the normal approximation (8), but this error vanishes as $n \to \infty$.

Unfortunately, Algorithm 2.1 is only computationally feasible for small K (usually $K \leq 3$) because the number of grid points needed for a given grid width d is proportional to d^{-K}, which grows exponentially with K.

2.3 Examples

We now provide some examples of Algorithm 2.1 in practice. In our examples we use $K = 2$ and $\bar{\Sigma} = I_K$, in which case the disturbances $\bar{m} - \mu$ are assumed to be spherical. To apply Algorithm 2.1, we need to choose a reasonable weighting π on the alternative against which we maximize power. We start by setting π to be a nearly flat (improper) distribution over H_a:

$$\pi \sim \mathcal{N}(0, V) \cap H_a, \tag{13}$$

where

$$V = \begin{pmatrix} 10^4 & 0 \\ 0 & 10^4 \end{pmatrix}.$$

The acceptance region for an approximate most powerful test of H_0 (5) against H_π (6) is shown by the dark region in Figure 1. Note that this region is well approximated by the union of a circle and the region bounded by two lines, as indicated in white in the figure.

Suppose instead that we have some prior belief that μ_1 and μ_2 are close to each other, which for example might be the case if we started with a conditional moment inequality model $\mathrm{E}[m(Y, \theta_0) \mid X = x] \leq 0$ and reduced it to an unconditional model using $\mu_1 = \mathrm{E}[m(Y, \theta_0) \mid X = x_1]$ and $\mu_2 = \mathrm{E}[m(Y, \theta_0) \mid X = x_2]$. Then we might choose to maximize power against the distribution (13) with V chosen to have higher weight for $\mu_1 \approx \mu_2$:

$$V = \begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix}. \tag{14}$$

The resulting approximately optimal test is shown by the dark region in Figure 2. The circular region is now replaced by roughly an elongated ellipse. Although the white ellipse and lines now look like a fairly bad approximation, we will discuss the test that accepts in the region bounded by the ellipse and two lines in Section 3 and show it has power close to the most powerful test.

3 A feasible test for moment inequalities

Algorithm 2.1 and other general algorithms for approximating most powerful tests are only computationally feasible for small K. In this section we present a faster algorithm that produces tests with power very close to the maximum power. We also show that these tests are exactly optimal within an intuitively reasonable class of tests.

The inspiration for the algorithm is the roughly elliptical shapes in Figures 1 and 1. To derive where these elliptical shapes come from, consider a test of the *simple* null hypothesis $\mu = 0$ against the alternative H_π (6) with

$$\pi \sim \mathcal{N}(0, V) \tag{15}$$

for some positive definite matrix V. Note that unlike in (13), this is not a true "alternative" to H_0 since it overlaps H_0 with probability greater than zero. However, we will use (15) here as an approximation to the actual alternative (13) because (15) yields a simple formula for the likelihood ratio and hence is easier to analyze. When we run simulations to compare the power of tests, we will always compute power against the actual alternative (13).

We continue to use the asymptotic normal approximation (8) for $\bar{m} - \mu$. Then under the null $\mu = 0$, the likelihood is proportional to $\exp\left(-\frac{1}{2}\bar{m}'\bar{\Sigma}^{-1}\bar{m}\right)$, and under the alternative, the likelihood is proportional to $\exp\left(-\frac{1}{2}\bar{m}'(V+\bar{\Sigma})^{-1}\bar{m}\right)$. The log likelihood ratio, ignoring an additive constant, is

$$\ell = \left(-\frac{1}{2}\bar{m}'\left(V+\bar{\Sigma}\right)^{-1}\bar{m}\right) - \left(-\frac{1}{2}\bar{m}'\bar{\Sigma}^{-1}\bar{m}\right) = \bar{m}'L\bar{m}, \tag{16}$$

where

$$L = \frac{1}{2}\left(\bar{\Sigma}^{-1} - \left(V+\bar{\Sigma}\right)^{-1}\right).$$

In order for the contours of the likelihood ratio to be ellipses, L must be positive semidefinite, which is a direct consequence of the following lemma given by Bhatia (2007):

Lemma 3.1 *Suppose that A and B are positive definite matrices and $A-B$ is positive semidefinite. Then $B^{-1} - A^{-1}$ is positive semidefinite.*

By the lemma, the acceptance region of a likelihood-ratio test at any critical value is an ellipse for $K = 2$, or an ellipsoid for $K = 3$. In the special case that V and $\bar{\Sigma}$ are both scalar multiples of the identity matrix, it is easy to see that L will be a scalar multiple of the identity matrix as well, and hence the acceptance region for $K = 2$ is a circle, which is what we see in Figure 1.

The likelihood-ratio test based on ℓ in (16) is the most powerful test against H_π that controls size for $\mu = 0$ by the Neyman-Pearson lemma, but it may not have correct size for other parameter values in our null hypothesis H_0 (5). This is why Figures 1 and 2 have larger acceptance regions than just the ellipse.

To attempt to create a test that controls size everywhere in H_0 (5), we constrain the test to always accept whenever a test on a subset of the constraints accepts and there is no "evidence" that the other constraints are violated, in a sense that we will formalize shortly. For example, for the case $K = 2$ and $\bar{\Sigma} = I_K$, if $\bar{m}_2 < 0$ we add in a test of the univariate hypothesis $\mu_1 \leq 0$, and if $\bar{m}_1 < 0$ we add in a test of the univariate hypothesis $\mu_2 \leq 0$. These two additional tests correspond to the regions to the left of the vertical white line and below the horizontal white line in Figures 1 and 2. Then, if we construct a likelihood-ratio test for the null hypothesis $\mu = 0$ under the constraint that it must accept in those regions, we add an ellipse to the acceptance region to get the test that accepts in the whole region enclosed by the ellipse and lines in 1 and 2. If this likelihood-ratio test has correct size α over all of H_0 (5), it is also a most powerful test under the constraint by Theorem 3.8.1 of Lehmann and Romano (2005).

This technique can be formalized and generalized recursively to higher dimensions. We construct a test of H_0 against any distribution π as follows:

Algorithm 3.1 *Algorithm for size-α test $\phi_K(\bar{m}; \alpha, \bar{\Sigma}, \pi)$ of H_0 (5) against alternative H_π (6) under the approximation $\bar{m} \sim \mathcal{N}(\mu, \bar{\Sigma})$:*

1. *If $K = 1$, $\phi_1(\bar{m}; \alpha, \bar{\Sigma}, \pi) = \mathbf{1}\left\{\Phi\left(\frac{\bar{m}}{\bar{\Sigma}^{1/2}}\right) > 1 - \alpha\right\}$.*

2. *If $K > 1$, then for each $k \in \{1, \ldots, K\}$ such that*

$$\bar{m}_k \leq \mathrm{E}\left[\bar{m}_k \mid \bar{m}_{-k}; \mu = 0\right] \tag{17}$$

and

$$\phi_1(\bar{m}_k; \alpha, \bar{\Sigma}_{k,k}, \pi_k) = 0, \tag{18}$$

compute $\phi_{K-1}(\bar{m}_{-k}; \alpha, \bar{\Sigma}_{-k,-k}, \pi_{-k})$, where $\bar{m}_{-k}$ denotes $\bar{m}$ with coordinate k excluded, $\bar{\Sigma}_{-k,-k}$ denotes $\bar{\Sigma}$ with row k and column k excluded, and π_{-k} is the marginal distribution of π over all coordinates except k. If any of these tests accepts the null hypothesis, return $\phi_K(\bar{m}; \alpha, \bar{\Sigma}, \pi) = 0$.

3. *Otherwise, compute the log likelihood ratio ℓ, evaluated at $\bar{m}$, for the test of the simple null $\mu = 0$ against the simple alternative $\mu \sim \pi$.*[6] *Compare ℓ to the critical value c obtained by running this algorithm on a large number of simulated draws of $\bar{m} \sim \mathcal{N}(0, \bar{\Sigma})$. Return $\phi_K(\bar{m}; \alpha, \bar{\Sigma}, \pi) = \mathbf{1}\{\ell > c\}$.*

The key step of the algorithm is step 2. Roughly, it states that if the test accepts the null if it would have accepted the null based on any $K - 1$ of the moment conditions and the other moment condition provides no additional evidence that the null is violated. Note that if $\bar{\Sigma}$ is diagonal, (17) and (18) simplify to $m_k \leq 0$. As a result of the recursive nature of step 2, the algorithm must compute up to $2^K - 1$ critical values. Later, we will present a faster approximation of the algorithm that only requires computing up to K critical values.

The acceptance region of Algorithm 3.1 for $\alpha = 0.05$, $\bar{\Sigma} = I_3$, and a flat alternative π over $\mathbb{R}^3$ is shown graphically in Figure 3. It is the union of a sphere (test for $K = 3$), three cylinders (tests for $K = 2$), and an area bounded by three planes (tests for $K = 1$).

We analyze Algorithm 3.1 by examining the actual size and power of tests produced by the algorithm.

[6] If π is of the form (15), then ℓ can be computed using (16).

3.1 Size control

We discuss the finite-sample size of the test $\phi_K(\cdot;\alpha,\bar{\Sigma},\pi)$ under the assumption that the asymptotic normal approximation (8) is correct. Showing that the test has correct size α in finite samples under this approximation implies that the size converges to α as $n \to \infty$ since the total variation distance between the approximate (8) and actual distributions of $\bar{m}-\mu$ goes to zero by Lemma 2.3. If that total variation distance goes to zero uniformly over μ then the convergence of size is uniform over μ. The need for uniform convergence has been emphasized recently by many authors, including Imbens and Manski (2004), because tests whose size do not converge uniformly may sometimes significantly overreject the null in finite samples.

Step 3 of Algorithm 3.1 guarantees that all tests produced by the algorithm have size α at $\mu = 0$. We present a theorem that gives conditions sufficient to prove that $\phi_K(\bar{m};\alpha,\bar{\Sigma},\pi)$ has correct size. Although the conditions are quite restrictive, we will next provide simulation evidence that Algorithm 3.1 controls size properly over a much wider range of parameter values than covered in the theorem.

Theorem 3.2 *Suppose that $\bar{\Sigma}$ is diagonal and that $\phi_K(\cdot;\alpha,\bar{\Sigma},\pi)$ satisfies the following condition: For any $\bar{m},\bar{m}' \in \mathbb{R}^K$ such that $0 \le \bar{m}'_k \le \bar{m}_k$ and $\bar{m}'_{-k} = \bar{m}_{-k}$ for some $k \in \{1,\ldots,K\}$,*

$$\phi_K(\bar{m};\alpha,\bar{\Sigma},\pi) \ge \phi_K(\bar{m}';\alpha,\bar{\Sigma},\pi). \tag{19}$$

Then the test $\phi_K(\cdot;\bar{\Sigma},\pi)$ produced by Algorithm 3.1 does indeed have size α, assuming that the critical values in step 3 are computed exactly.

The theorem is proven in the Appendix.

The condition (19) requires a particular type of convexity of the acceptance region of $\phi_K(\cdot;\alpha,\bar{\Sigma},\pi)$. Note that if $\bar{\Sigma}$ is diagonal and π is an (improper) flat distribution over $\mathbb{R}^K$, or π has the form (15) with diagonal V, then the acceptance region of $\phi_K(\cdot;\alpha,\bar{\Sigma},\pi)$ must satisfy the condition (19) since it is a union of spherical and cylindrical regions centered at the origin and along the axes; see the region outlined in white in Figure 1 for an example. Even in Figure 2, in which V (14) is not close to diagonal, (19) is satisfied. However, if the ellipse becomes a bit more slanted, as will be the case if the correlation in V is raised, then (19) will no longer be satisfied, but simulations show that $\phi_K(\cdot;\alpha,\bar{\Sigma},\pi)$ still has correct size.

As an example, we plot the size of the test along one of the boundaries of H_0 for $\alpha = 0.05$, $\bar{\Sigma} = I_2$, and $\pi \sim \mathcal{N}(0,V)$ with

$$V = \begin{pmatrix} 1 & \rho_\pi \\ \rho_\pi & 1 \end{pmatrix}.$$

In Figure 4, $\rho_\pi = 0$, and $\phi_2(\cdot;\alpha,\bar{\Sigma},\pi)$ satisfies the conditions of Theorem 3.2, so the test has size at most α everywhere as expected. In Figure 5, $\rho_\pi = 0.9$, and the test still has correct size despite that (19) is not satisfied. In fact, the size is far less than

α in many places. Note, though, that non-similarity does not imply that the test has suboptimal power; for many problems no test exists that is similar.

Next, we try varying $\bar{\Sigma}$. Here, the actual size of the test can be greater than α, but only for very large negative correlations in $\bar{\Sigma}$. We use $\pi \sim \mathcal{N}(0, I_2)$, and suppose

$$\bar{\Sigma} = \begin{pmatrix} 1 & \rho_m \\ \rho_m & 1 \end{pmatrix}.$$

If $\rho_m = -0.9$, then the test still has valid size as shown in Figure 6. However, for $\rho_m = -0.99$, then the test has size 0.064 for nominal size $\alpha = 0.05$, as shown in Figure 7. In the worst case, $\rho_m = -1$, the actual size is 2α. Unfortunately, the case $\rho_m = -1$ occurs for some applications of interest, such as regressions with interval outcomes, in which the error term in one inequality is equal to the negative of the error term in the other inequality.

Even in such cases, the test has correct size at most μ in H_0, and size converges *pointwise* to α everywhere in H_0, which is stated in the theorem below and proven in the Appendix.

Theorem 3.3 *Under Assumption 1, for any $\mu^* \in H_0$ and distribution π on H_a,*

$$\lim_{n\to\infty} \mathrm{E}\left[\phi_K(\cdot; \alpha, \bar{\Sigma}, \pi) \mid \mu = \mu^*\right] \leq \alpha,$$

with equality if μ^* is on the boundary of H_0 (i.e. $\mu_k^* = 0$ for some k), where $\bar{\Sigma}$ is given by (8).

We will provide evidence of finite-sample size control for larger K in Section 4.2.

3.2 Power

For small K, we can directly compare the power of tests generated by Algorithm 3.1 with the approximately optimal tests constructed in Section 2. In the example of Figure 1, in which $\alpha = 0.05$, $\bar{\Sigma} = I_2$, and π is a nearly flat distribution given by (13), the two tests look very similar, and indeed, their powers are very similar. Algorithm 3.1 has power 0.4585 over the region shown in the figure, compared to 0.4591 for the approximately optimal test.

Figure 8 compares the acceptance region of Algorithm 3.1 with the acceptance regions of other tests in the literature. Note that the acceptance region for our test narrows away from the origin similarly to the generalized moment selection approach of Andrews and Soares (2007).

For the example of Figure 2, we conduct the same analysis for π given by (14), which assigns strong correlation to μ_1 and μ_2. While the two tests appear graphically to be quite different far away from the origin, there is little weight on the alternative in these regions, and hence the powers are similar: Algorithm 3.1 has power 0.2625, compared to 0.2701 for the approximately optimal test.

In Section 4.2, we will compare the power of Algorithm 3.1 to other tests for much larger K.

While we can only provide simulation evidence that tests generated by Algorithm 3.1 come close to the maximum power against π, there is a different sense in which the tests are *exactly* optimal. Whenever a test produced by Algorithm 3.1 controls size correctly, it is the most powerful within a reasonable class of tests, as expressed in the following theorem:

Theorem 3.4 *For a given $\bar{\Sigma}$ and π, if each member of the family of tests $\{\phi_k : k \in \{1, \ldots, K\}\}$ produced by Algorithm 3.1 does indeed have size α, then each of the tests maximizes power, subject to the following conditions:*

1. *When $k = 1$, $\phi_1(\bar{m}; \alpha, \bar{\Sigma}, \pi)$ has the form $\mathbf{1}\{\bar{m} > c\}$ for some $c \in \mathbb{R}$.*
2. *When $k \geq 2$, $\phi_k(Y) = 0$ if there exists $k^* \in \{1, \ldots, k\}$ such that $\bar{m}_k \leq \mathrm{E}[\bar{m}_k \mid \bar{m}_{-k}; \mu = 0]$ and $\phi_1(\bar{m}_k; \alpha, \bar{\Sigma}_{k,k}, \pi_k) = 0$ and $\phi_{k-1}(\bar{m}_{-k^*}; \alpha, \bar{\Sigma}_{-k^*,-k^*}, \pi_{-k^*}) = 0$*

Proof. The proof of the theorem is by induction on k.

For $k = 1$, a test is most powerful in the class given by condition 1 as long as the critical value c is chosen to yield exact size α. This is ensured by Step 1 of Algorithm 3.1.

For $k \geq 2$, assume ϕ_{k-1} satisfies the theorem. Step 2 of the algorithm ensures that ϕ_k satisfies condition 2. Then by the construction in Step 3 of the algorithm and the Neyman-Pearson Lemma, ϕ_k is the most powerful test satisfying condition 2 for the simple null $\mu = 0$. But since ϕ_k has size α across all of H_0, by Theorem 3.8.1 of Lehmann and Romano (2005) it is a most powerful test for H_0 as well. ■

Condition 2 of Theorem 3.4 is closely linked to Step 2 of Algorithm 3.1 and is quite intuitive: It says that if we would not reject H_0 based on a $k-1$ of the moment conditions, then we still do not reject H_0 after checking one additional moment if that moment condition provides no additional evidence that the null is violated.

3.3 A faster algorithm

Algorithm 3.1 may require the computation of up to $2^K - 1$ critical values, one for each subset of the moment conditions, and therefore it may be quite slow. The following algorithm produces the same results for the special case that $\bar{\Sigma}$ and V (15) are both symmetric with respect to all coordinates (as is the case if they are multiples of I_K), but only needs K critical values need to be computed.

Algorithm 3.2 *Faster algorithm for size-α test $\phi_K(\bar{m}; \alpha, \bar{\Sigma}, \pi)$ of H_0 (5) against alternative H_π (6) under the approximation $\bar{m} \sim \mathcal{N}(\mu, \bar{\Sigma})$:*

1. *If $K = 1$, $\phi_1(\bar{m}; \alpha, \bar{\Sigma}, \pi) = \mathbf{1}\left\{\Phi\left(\frac{\bar{m}}{\bar{\Sigma}^{1/2}}\right) > 1 - \alpha\right\}$.*

2. *If $K > 1$, among all $k \in \{1, \ldots, K\}$ that satisfy (17) and (18), choose $\hat{k}$ to minimize the log likelihood ratio, evaluated at $\bar{m}_{-\hat{k}}$, for the test of the simple null $\mu_{-\hat{k}} = 0$ against the simple alternative $\mu \sim \pi_{-K}$. If $\hat{k}$ exists, compute $\phi_{K-1}(\bar{m}_{-\hat{k}}; \alpha, \bar{\Sigma}_{-K,-K}, \pi_{-K})$. If this test accepts the null hypothesis, return $\phi_K(\bar{m}; \alpha, \bar{\Sigma}, \pi) = 0$.*

3. *Otherwise, compute the log likelihood ratio ℓ, evaluated at $\bar{m}$, for the test of the simple null $\mu = 0$ against the simple alternative $\mu \sim \pi$. Compare ℓ to the critical value c obtained by running this algorithm on a large number of simulated draws of $\bar{m} \sim \mathcal{N}(0, \bar{\Sigma})$. Return $\phi_K(\bar{m}; \alpha, \bar{\Sigma}, \pi) = \mathbf{1}\{\ell > c\}$.*

There are two shortcuts involved in Algorithm 3.2. First, in the recursive steps, we use the distributions $\bar{\Sigma}_{-K,-K}$ and π_{-K} rather than $\bar{\Sigma}_{-\hat{k},-\hat{k}}$ and $\pi_{-\hat{k}}$. That is, we only have to consider one pair of distributions for each K, and hence we only need to compute K critical values. If $\bar{\Sigma}$ and V are both symmetric with respect to all coordinates, then this shortcut is justified. Second, the algorithm makes only one recursive call in Step 2, so it is "greedy" in the sense that it does not check all subsets of the constraints but instead eliminates constraints one-by-one without backtracking. If $\bar{\Sigma}$ and V are both symmetric with respect to all coordinates, then by (16) the log likelihood is symmetric with respect to all coordinates of $\bar{m}$, so removing the most negative elements of $\bar{m}$ one-by-one is guaranteed to minimize the log-likelihood at each step subject to (17) and (18).

Algorithm 3.2 can be used for speed purposes even if $\bar{\Sigma}$ and V are not multiples of I_K, but one should run simulations to check that the algorithm has valid size for the chosen parameters.

4 An application to conditional moment inequalities

We run large-scale simulations of Algorithm 3.2 in the context of testing the conditional moment inequality model

$$\mu(x) \ \leq 0 \text{ for all } x \in \mathcal{X}, \tag{20}$$

where $\mu(x) = \mathrm{E}\left[m(X, Y, \theta) \mid X = x\right]$ at a given parameter value θ.

We observe n i.i.d. draws $\{(X_i, Y_i)\}_{i=1}^n$ of the random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$. We will deal with the case where μ is a scalar-valued function, so that we are testing one conditional moment inequality. There are several ways that we might adapt this model to the unconditional framework of Sections 2 and 3. One possibility is to partition the n samples into $K = \sqrt{n}$groups $\mathcal{I}_1, \ldots, \mathcal{I}_K$ of $\sqrt{n}$ samples, where the clustering is done based on X_i.[7] Grouping based on X_i makes sense if we

[7] More generally, the number of groups might be given by any function $f(n)$ such that $\lim_{n\to\infty} f(n) = \infty$ and $\lim_{n\to\infty} \frac{n}{f(n)} = \infty$.

believe that the function $\mu(x)$ is smooth and we want our test to have power against local violations of (20). We then can use Algorithm 3.2 for the K inequalities $\mathrm{E}\left[m(X_i, Y_i, \theta) \mid i \in \mathcal{I}_k\right] \leq 0$ for $k \in \{1, \ldots, K\}$, and we know that $\bar{\Sigma}$ will be diagonal due to the i.i.d. assumption.

4.1 A smoothness prior on the alternative

Another possibility is to create one inequality for each distinct observed value of X. If X is a continuous random variable, then with probability one this will yield one inequality per observation. The advantage of this approach is that we are not throwing out any information by averaging observations, but the disadvantage is that we can no longer use a central-limit theorem as in (7) if X is continuously distributed because we will have only one observation per inequality, even as $n \to \infty$. However, if we knew the distribution of $m(X_i, Y_i, \theta) - \mu(X_i)$, then we could still apply Algorithm 3.1.

Define the vectors $\mu = (\mu(X_i))_{i=1}^n$ and $\bar{m} = (m(X_i, Y_i, \theta))_{i=1}^n$. Assume that we know that $\bar{m} \sim \mathcal{N}(\mu, \bar{\Sigma})$ where $\bar{\Sigma} = I_K$. Intuitively, one would usually expect the function $\mu(x)$ to be smooth, and hence we choose our prior π to favor smoothness.[8] One way to favor smoothness is impose a high correlation of μ_i and μ_j when X_i is close to X_j.

Specifically, we choose

$$H_\pi : \mu \sim \mathcal{N}(0, V),$$

where for all $i, j \in \{1, \ldots, n\}$,

$$V_{i,j} = \sigma^2 \rho^{d_{i,j}} \tag{21}$$

for some constants $\rho \in [0, 1]$ and $\sigma^2 > 0$ of our choice. While changing the parameters of the alternative distribution does not affect the size of a resulting test, it does alter the power against various alternatives.

4.2 Simulations and comparisons to other tests

We compare the power and size of various tests of conditional moment inequalities. We choose $n = 100$, and run tests of $H_0 : \mathrm{E}[m(Y, \theta) \mid X = x] \leq 0$ for all x, where

$$m(Y, \theta) = y - \theta.$$

In all simulations, X is drawn from the uniform distribution on $[0, 1]$, and $Y \mid X \sim \mathcal{N}(f(X), 1)$ for some function $f(X) = \mathrm{E}[Y \mid X = x]$. We try out a few different functions for the true $f(X)$, and the results are shown in Tables 1-4 below.

[8] When we split the sample into larger groups, it was not necessary to choose π to favor smoothness because we locally smoothed the data by averaging the samples within each group.

The first test considered is the test of Müller and Watson (2008), with the sample broken into $\sqrt{n}$ groups based on the ordering of X_i. Rosen's test statistic is

$$\sum_{k=1}^{K} \left(\max\left\{\bar{m}_k, 0\right\}\right)^2,$$

and the critical value is computed at the worst-case point $\mu = 0$. The GMS method of Andrews and Soares (2007) improves the power of Rosen's test by decreasing the critical value whenever some inequalities are determined to be satisfied with very high probability. We also run Algorithm 3.2 with the samples grouped in this way against a flat distribution π over the alternative.

Among methods that are designed particularly for conditional moment inequalities, we apply Kim's (2008) test (4) using the worst-case critical value, and we run Algorithm 3.2 with one inequality per data point as described in Section 4.1 using the alternative π given by (21) with $\sigma^2 = 4$ and three different values of ρ: $\rho = 0$, $\rho = \frac{1}{2}$, and $\rho = 1$. It is important to note the tests of Section 4.1 "cheat" relative to the other tests because they assume knowledge of the distribution $\mathcal{N}(0, 1)$ of $m(Y, \theta) - \mathrm{E}\left[m(Y, \theta)\right]$.

In Table 1, we set $f(x) = 1$ for all x. In this case the null hypothesis is true for $\theta = 1$ and false for $\theta < 1$. Since $\mathrm{E}[Y \mid X]$ is constant, an unconditional test on $Y - \theta$ actually works best here. In fact, Algorithm 3.2 with the alternative (21) with $\rho = 1$ assumes that $\mu(x)$ is constant over x and hence produces a test that is nearly identical to the unconditional test; it performs well in this case. Using the alternative $\rho = 0$ always works poorly because then the test is not tailored for smooth $\mu(x)$. All of the tests using grouped data significantly overreject for $\theta = 1$, when H_0 is true. This is because with the grouping, we effectively are using $n = 10$, which is too small for a normal approximation to work. Since all three tests overreject equally, the power of these tests can be compared with each other but not with the other tests.

The test with $\rho = \frac{1}{2}$ outperforms all other correctly-sized tests in Tables 2 and 3. In these examples, $f(x)$ varies with x so that some inequalities are closer to binding than others, and only our test and the GMS test of Müller and Watson (2008) are designed to have good power in such situations. However, the variation of $f(x)$ with x is insufficient for GMS to detect it for individual observations, so GMS does not work well.

In Table 4, the variation of $f(x)$ with x is drastic enough that GMS does detect it well and performs similarly to Algorithm 3.2.

5 Conclusions and future work

We have used a general algorithm to approximate the most powerful test for a set of moment inequalities against a chosen weighting over the alternative hypothesis. Such general algorithms are only computationally feasible for small-dimensional problems, but we have developed a faster test that has power close to the maximum power in low dimensions but also generalizes to high dimensions as well. The test's size converges to the nominal size uniformly for many models and pointwise for the rest. It is shown to be the most powerful test within a reasonable class of tests whenever it has correct size. The test can be adapted to conditional moment inequality models, and simulations for those models show that in most cases our algorithms produce tests that are more powerful than other tests in the literature.

One direction for future work is to find a way to modify Algorithm 3.1 so that it has correct size given $\bar{\Sigma}$, even in the difficult cases in which $\bar{\Sigma}$ has large negative correlation. Such a test would have asymptotically correct size uniformly.

Another extension is to derive approximately most powerful tests for a *subset* of the parameters θ, treating the other parameters as nuisance parameters. The algorithms in this paper can be used as the first step in an procedure for testing a subset of the parameters, but one must design the entire procedure to be approximately most powerful.

A Appendix

A.1 The dual linear program

Algorithm 2.1 has a duality connection to algorithms for optimal tests such as that of Müller and Watson (2008) that approximate the least favorable prior on the null and use a likelihood-ratio test of the simple null based on the least favorable prior. We show here that the dual problem of (9) is in fact a search for the least favorable prior on the null. Krafft and Witting (1967) derive a similar duality result for an algorithm for finding maximin-power tests.

The dual problem of (9) is

$$\begin{aligned}
\min_{\{\lambda_j\},\{g_s\}} \quad & \alpha \sum_{j=1}^{J} \lambda_j + \sum_{s \in S} g_s \qquad (22)\\
\text{s.t.} \quad & \sum_{j=1}^{J} \nu_{j,s} \lambda_j + g_s \geq \pi_s \text{ for all grid squares } s\\
& \lambda_j \geq 0 \text{ for all } j \in \{1, \ldots, J\}\,;\; g_s \geq 0 \text{ for all } s \in S.
\end{aligned}$$

Let $c = \sum_{j=1}^{J} \lambda_j$, and for each j define $\tilde{\lambda}_j = \lambda_j / c$, so that $\sum_{j=1}^{J} \tilde{\lambda}_j = 1$. We will see that $\left\{ \tilde{\lambda}_j \right\}$ can be interpreted as a prior over the null points $\{\mu^j\}$, and c can be interpreted as a critical value of a likelihood-ratio test.

In the optimal solution, we will have $g_s = 0$ when $c \sum_{j=1}^{J} \nu_{j,s} \tilde{\lambda}_j \geq \pi_s$; this is equivalent to the ratio of likelihood under the alternative to likelihood under the null with prior λ being at most c, or the acceptance region of a likelihood-ratio test of the simple null $\tilde{\lambda}$ against the simple alternative π. In the rejection region, $g_s = \pi_s - \sum_{j=1}^{J} \nu_{j,s} c\tilde{\lambda}_j$. Then we can write the objective function as

$$\begin{aligned} & \alpha c + \sum_{s \text{ rejected}} (\pi_s - \sum_{j=1}^{J} \nu_{j,s} c\tilde{\lambda}_j) \\ = \quad & \alpha c + \sum_{s \text{ rejected}} \pi_s - c \sum_{j=1}^{J} \left(\tilde{\lambda}_\mu \sum_{s \text{ rejected}} \nu_{j,s} \right) \\ = \quad & \alpha c + \sum_{s \text{ rejected}} \pi_s - c(\text{size of LR test of } \tilde{\lambda} \text{ against } \pi). \end{aligned} \tag{23}$$

We claim that c is chosen so that the size of the likelihood-ratio test in (23) is α. For any particular $\tilde{\lambda}$, the optimal value of c is determined by

$$\min_c \alpha c + \sum_{s \in S} \left(\mathbf{1} \left\{ \sum_{j=1}^{J} \nu_{j,s} c\tilde{\lambda}_j < \pi_s \right\} \left(\pi_s - \nu_{j,s} c\tilde{\lambda}_j \right) \right).$$

The first-order condition is (we actually need to be more careful and use left and right derivatives at discontinuities of the indicator function)

$$\begin{aligned} \alpha + \sum_{s \in S} \left(\mathbf{1}\left\{\text{reject}\right\} \left(-\nu_{j,s}\tilde{\lambda}_j \right) \right) &= 0 \\ \alpha - (\text{size of LR test of } \tilde{\lambda} \text{ against } \pi) &= 0. \end{aligned}$$

Plugging this into (23), we get that the objective function of (22) is equal to

$$\sum_{s \text{ rejected}} \pi_s = \text{power of LR test of } \tilde{\lambda} \text{ against } \pi.$$

That is, the dual program finds the prior $\tilde{\lambda}$ on the null such that the size-α LR test of the simple null $\tilde{\lambda}$ against the simple alternative π has *least* power, i.e. it finds the *least favorable prior* on the null.

A.2 Approximating the least favorable prior using importance sampling

As shown in Appendix A.1, Algorithm 2.1 approximates the least favorable distribution on the null. However, Algorithm 2.1 requires creating a grid for $\mathcal{Y}$, and the number of grid squares required to attain a given level of precision grows exponentially with K. Other algorithms exist that approximate the least favorable prior, that do not divide $\mathcal{Y}$ into a grid but instead sample from $\mathcal{Y}$, e.g. Müller and Watson (2008). Although these algorithms suffer from the curse of dimensionality as well because the number of points in a least favorable prior also grows exponentially with K, it is less severe.

We present here an algorithm that uses importance sampling:

Algorithm A.1 *Algorithm for computing least favorable prior over null points* $\{\mu^j\}_{j=1}^J$ *against alternative* π*:*

1. Let $\pi_{\bar{m}}$ be the density on $\bar{m}$ obtained by convolving the distribution π of μ with the distribution (8) of $\bar{m} - \mu$. Draw a large number K of samples $\left\{\bar{m}^k\right\}_{k=1}^K$ from $\pi_{\bar{m}}$. This may be done using a Markov chain Monte Carlo method.

2. For each j, k, compute

$$\nu_{j,k} = \frac{\exp\left(-\frac{n}{2}(\bar{m}^k - \mu^j)'\hat{\Sigma}^{-1}(\bar{m}^k - \mu^j)\right)}{T_j \pi_{\bar{m}}(\bar{m}^k)}$$

   where

$$T_j = \sum_{k=1}^K \frac{\exp\left(-\frac{n}{2}(\bar{m}^k - \mu^j)'\hat{\Sigma}^{-1}(\bar{m}^k - \mu^j)\right)}{\pi_{\bar{m}}(\bar{m}^k)}$$

   This is an importance-sampling approximation to the distribution of $\bar{m}$ when μ^j is the true value of μ.

3. Solve the linear program

$$\begin{aligned} \max_{\{\phi_k\}} \quad & \frac{1}{K}\sum_{k=1}^K \phi_k \\ \text{s.t.} \quad & \sum_{k=1}^K \nu_{j,k}\phi_k \leq \alpha \text{ for all } j \in \{1, \ldots, J\} \\ & 0 \leq \phi_k \leq 1 \text{ for all } k \in \{1, \ldots, K\}. \end{aligned}$$

4. The resulting Lagrange multipliers on the J size constraints are the least favorable prior over the null points $\{\mu^j\}_{j=1}^J$, with respect to the importance-sampling approximation of the null distributions and alternative distribution, by the duality argument of Appendix A.1.

To get an even better approximation, one can run this algorithm several times, adjusting the set $\{\mu^j\}$ after each iteration as Müller and Watson (2008) do with their algorithm.

A.3 Proof of Theorem 3.2

Without loss of generality, assume that $\bar{\Sigma} = I_K$, since this can be achieved via a rescaling of $\bar{\Sigma}$, $\bar{m}$, and π given that $\bar{\Sigma}$ is diagonal. We prove the theorem by induction on K.

For $K = 1$, $\phi_1(\bar{m}; \alpha, \bar{\Sigma}, \pi) = \mathbf{1}\{\Phi(\bar{m}) > 1 - \alpha\}$ has size α at $\mu = 0$ because $\Pr[\bar{m} > z \mid \mu = 0] = \Phi(z)$. Since $\Pr[\bar{m} > z \mid \mu]$ is increasing with μ, it has size at most α for all $\mu < 0$ as well.

For $K > 1$, assume that for each $k \in \{1, \ldots, K\}$,

$$\phi_{K-1}(\cdot; \pi_{-k}) \text{ has size } \alpha. \tag{24}$$

Consider any $\mu \leq 0$. We perform another induction on the smallest k such that $\mu_k < 0$.

If there is no such k, then $\mu_k = 0$ for all k. In that case, the test has size α by step 3 of the algorithm.

Otherwise, find the smallest k such that $\mu_k < 0$. We assume the following inductive hypothesis: For all μ^* such that $\mu^*_j \leq 0$ for all $j \in \{1, \ldots, k-1\}$ and $\mu^*_j = 0$ for all $j \in \{k, \ldots, K\}$, $\phi_K(\cdot; \alpha, \bar{\Sigma}, \pi)$ has size at most α at μ^*.

Define μ^* such that $\mu^*_k = 0$ and $\mu^*_j = \mu_j$ for all $j \neq k$. For each $z \in \mathbb{R}$, define

$$r(z) = \mathrm{E}\left[\phi_K(\bar{m}; \alpha, \bar{\Sigma}, \pi) \mid \bar{m}_k = z \text{ and } \bar{m}_{-k} \sim \mathcal{N}(\mu_{-k}, I_{K-1})\right].$$

Then the size at μ^* is given by

$$\pi(\mu^*) = \int_{-\infty}^{\infty} r(z)\Phi'(z)dz, \tag{25}$$

and the size at μ is

$$\pi(\mu) = \int_{-\infty}^{\infty} r(z)\Phi'(z - \mu_k)dz. \tag{26}$$

We know that $\pi(\mu^*) \leq \alpha$ by the inductive hypothesis, and our goal is to show that $\pi(\mu) \leq \alpha$. To get there, we establish two facts about $r(\cdot)$.

First, by step 2 of Algorithm 3.1, $\phi_K(\bar{m}; \alpha, I_K, \pi) = 0$ if $\bar{m}_k \leq 0$ and $\phi_{K-1}(\bar{m}; \alpha, I_{K-1}, \pi) = 0$, so for all $z \leq 0$,

$$\begin{aligned} r(z) &\leq \mathrm{E}\left[\phi_{K-1}(\bar{m}; \alpha, I_{K-1}, \pi) = 0 \mid \bar{m} \sim \mathcal{N}(\mu_{-k}, I_{K-1})\right] \\ &= \text{size of } \phi_{K-1}(\cdot; \alpha, I_{K-1}, \pi) \text{ at } \mu_{-k} \\ &\leq \alpha \end{aligned} \tag{27}$$

by (24).

Second, we want to establish that $r(\cdot)$ is increasing on the interval $[0, \infty)$. This follows immediately from the condition (19) given in the statement of the theorem.

Since $r(z) \leq \alpha$ for all $z \leq 0$ by (27), and $r(\cdot)$ is increasing on $[0, \infty)$, there exists $\hat{z} \in \mathbb{R}$ such that $r(z) \leq \alpha$ for all $z \leq \hat{z}$, and $r(z) > \alpha$ for all $z > \hat{z}$. Now, we rewrite (26) as

$$\pi(\mu) - \alpha = \int_{-\infty}^{\hat{z}} (r(z) - \alpha)\Phi'(z - \mu_k)dz + \int_{\hat{z}}^{\infty} (r(z) - \alpha)\Phi'(z - \mu_k)dz.$$

Let

$$c = \frac{\Phi'(\hat{z} - \mu_k)}{\Phi'(\hat{z})}.$$

It is easy to verify that $\frac{\Phi'(z-\mu_k)}{\Phi'(z)}$ is a decreasing function of z. Therefore,

$$\begin{aligned} \pi(\mu) - \alpha &\leq \int_{-\infty}^{\hat{z}} c(r(z) - \alpha)\Phi'(z)dz + \int_{\hat{z}}^{\infty} c(r(z) - \alpha)\Phi'(z)dz \\ &= c\left(\pi(\mu^*) - \alpha\right). \end{aligned}$$

By the inductive hypothesis, $\pi(\mu^*) \leq \alpha$. It follows that $\pi(\mu) \leq \alpha$.

A.4 Proof of Theorem 3.3

Let $\mathcal{K}_0 = \{k \in \{1, \ldots, K\} : \mu_k^* = 0\}$, and let $\mathcal{K}_1 = \{k \in \{1, \ldots, K\} : \mu_k^* < 0\}$. Suppose that $\mathcal{K}_0 \neq \varnothing$, so that μ^* is on the boundary of H_0. Algorithm 3.1 includes a test of the lower-dimensional hypothesis $\mu_k \leq 0$ for all $k \in \mathcal{K}_0$; this test is used whenever $\phi_1(\bar{m}_k; \alpha, \bar{\Sigma}_{k,k}, \pi_k) = 0$ for all $k \in \mathcal{K}_1$, which occurs with probability approaching 1 as $n \to \infty$ since $\bar{\Sigma} \xrightarrow{p} 0$, and if there exists an ordering (k_i) of $\mathcal{K}_1$ such that

$$\bar{m}_{k_i} \leq \mathrm{E}\left[\bar{m}_{k_i} \mid \left\{\bar{m}_{k_j} : j > i\right\} \cup \left\{\bar{m}_k : k \in \mathcal{K}_0\right\}; \mu = 0\right]$$

for all i. To obtain this, we simply order $\mathcal{K}_1$ in increasing order of $\bar{m}_k/\bar{\Sigma}_{k,k}$. This works because $\lim_{n\to\infty} \Pr\left[\bar{m}_k \leq 0 \text{ for all } k \in \mathcal{K}_1\right] = 1$ since $\bar{\Sigma} \xrightarrow{p} 0$, and hence the order is decreasing in unlikeliness. Therefore, Algorithm 3.1 includes a test of the lower-dimensional hypothesis $\mu_k \leq 0$ for all $k \in \mathcal{K}_0$ with probability 1. By construction, this test will have size α at $\mu_k = 0$ for all $k \in \mathcal{K}_0$ under the normal approximation (8). Since the acceptance region of the entire algorithm includes the acceptance region of this test, the size of $\phi_K(\cdot; \alpha, \bar{\Sigma}, \pi)$ at μ^* is at most α under (8) with probability approaching 1. By Lemma 2.3, the total variation distance between the approximation (8) for the distribution of $\bar{m}$ and the true distribution goes to 0 as $n \to \infty$, so any error in the size due to the normal approximation goes to zero.

We still need to consider the case $\mathcal{K}_0 = \varnothing$. In this case, by the same reasoning as above, with probability approaching 1 as $n \to \infty$, Algorithm 3.1 will consider a test of a single-dimensional hypothesis. Since $\bar{\Sigma} \xrightarrow{p} 0$, step 1 of the algorithm accepts all single-dimensional hypotheses with probability approaching 1 as $n \to \infty$. Therefore, $\lim_{n\to\infty} \mathrm{E}\left[\phi_K(\cdot; \alpha, \bar{\Sigma}, \pi) \mid \mu = \mu^*\right] = 0$.

References

Andrews, D. W. K., and W. Ploberger (1994): "Optimal Tests when a Nuisance Parameter is Present Only Under the Alternative," *Econometrica*, 62(6), 1383–1414.

Andrews, D. W. K., and G. Soares (2007): "Inference for Parameters Defined by Moment Inequalities using Generalized Moment Selection," *Cowles Foundation Discussion Paper*, 1631.

Bhatia, R. (2007): *Positive Definite Matrices.* Princeton University Press, Princeton.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243–1284.

Chiburis, R. C. (2008): "Semiparametric Bounds on Treatment Effects," *Working paper, Princeton University.*

Domínguez, M. A., and I. N. Lobato (2004): "Consistent Estimation of Models Defined by Conditional Moment Restrictions," *Econometrica*, 72(5), 1601–1615.

Imbens, G. W., and C. F. Manski (2004): "Confidence Intervals of Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.

Khan, S., and E. Tamer (2006): "Inference on Randomly Censored Regression Models Using Conditional Moment Inequalities," *Working paper, Northwestern University.*

Kim, K. I. (2008): "Set Estimation and Inference with Models Characterized by Conditional Moment Inequalities," *Working paper, University of Minnesota.*

Krafft, O., and H. Witting (1967): "Optimale Tests und Ungünstigste Verteilungen," *Probability Theory and Related Fields*, 7(4), 289–302.

Lehmann, E. L., and J. P. Romano (2005): *Testing Statistical Hypotheses.* Springer, New York, 3rd edn.

Manski, C. F., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

Müller, U. K., and M. W. Watson (2008): "Low-Frequency Robust Cointegration Testing," *Working Paper, Princeton University.*

Rosen, A. M. (2007): "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities," *Working paper, University College London.*

Table 1: Rejection probabilities, in percentages, for the data-generating process $E[Y \mid X = x] = 1$ for all $x \in [0, 1]$.

θ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Rosen (grouped)	100	100	100	100	100	100	97	85	59	32	13
GMS (grouped)	100	100	100	100	100	100	97	85	59	33	14
Alg. 3.2 (grouped)	100	100	100	100	100	99	94	78	53	30	14
Kim worst-case	100	100	100	100	100	100	96	85	56	23	5
Sec. 4.1, $\rho = 1$	100	100	100	100	100	100	99	91	64	26	6
Sec. 4.1, $\rho = \frac{1}{2}$	100	100	100	100	100	100	96	82	50	20	4
Sec. 4.1, $\rho = 0$	100	100	100	99	94	81	59	40	23	12	4

Table 2: Rejection probabilities, in percentages, for the data-generating process $E[Y \mid X = x] = x$ for all $x \in [0, 1]$.

θ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Rosen (grouped)	100	99	95	84	63	39	22	8	3	1	0
GMS (grouped)	100	99	95	85	66	44	26	13	5	2	1
Alg. 3.2 (grouped)	100	98	93	82	65	44	28	16	9	5	2
Kim worst-case	100	98	90	68	38	14	24	1	0	0	0
Sec. 4.1, $\rho = 1$	100	99	90	64	27	6	0	0	0	0	0
Sec. 4.1, $\rho = \frac{1}{2}$	100	99	94	80	57	30	14	4	1	0	0
Sec. 4.1, $\rho = 0$	90	77	59	42	24	12	4	1	0	0	0

Table 3: Rejection probabilities, in percentages, for the data-generating process $E[Y \mid X = x] = \mathbf{1}\left\{x > \frac{1}{2}\right\}$ for all $x \in [0, 1]$.

θ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Rosen (grouped)	100	100	99	97	90	76	59	36	18	7	2
GMS (grouped)	100	100	99	98	92	81	66	50	30	16	7
Alg. 3.2 (grouped)	100	100	99	97	93	83	71	53	34	18	9
Kim worst-case	100	100	96	86	71	47	22	6	1	0	0
Sec. 4.1, $\rho = 1$	100	98	88	63	28	7	1	0	0	0	0
Sec. 4.1, $\rho = \frac{1}{2}$	100	100	100	98	90	76	54	31	13	3	1
Sec. 4.1, $\rho = 0$	98	94	84	70	53	33	15	4	1	0	0

Table 4: Rejection probabilities, in percentages, for the data-generating process $E[Y \mid X = x] = 1 - 10\left(\mathbf{1}\left\{x > \frac{1}{2}\right\}\right)$ for all $x \in [0, 1]$.

θ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Rosen (grouped)	100	100	98	96	88	74	56	34	17	8	4
GMS (grouped)	100	100	100	98	96	90	78	60	38	20	9
Alg. 3.2 (grouped)	100	100	100	98	96	89	75	56	35	19	9
Kim worst-case	87	77	66	50	32	17	7	2	1	0	0
Sec. 4.1, $\rho = 1$	0	0	0	0	0	0	0	0	0	0	0
Sec. 4.1, $\rho = \frac{1}{2}$	100	100	100	100	98	91	78	56	30	14	5
Sec. 4.1, $\rho = 0$	100	99	98	93	83	68	46	30	16	8	3

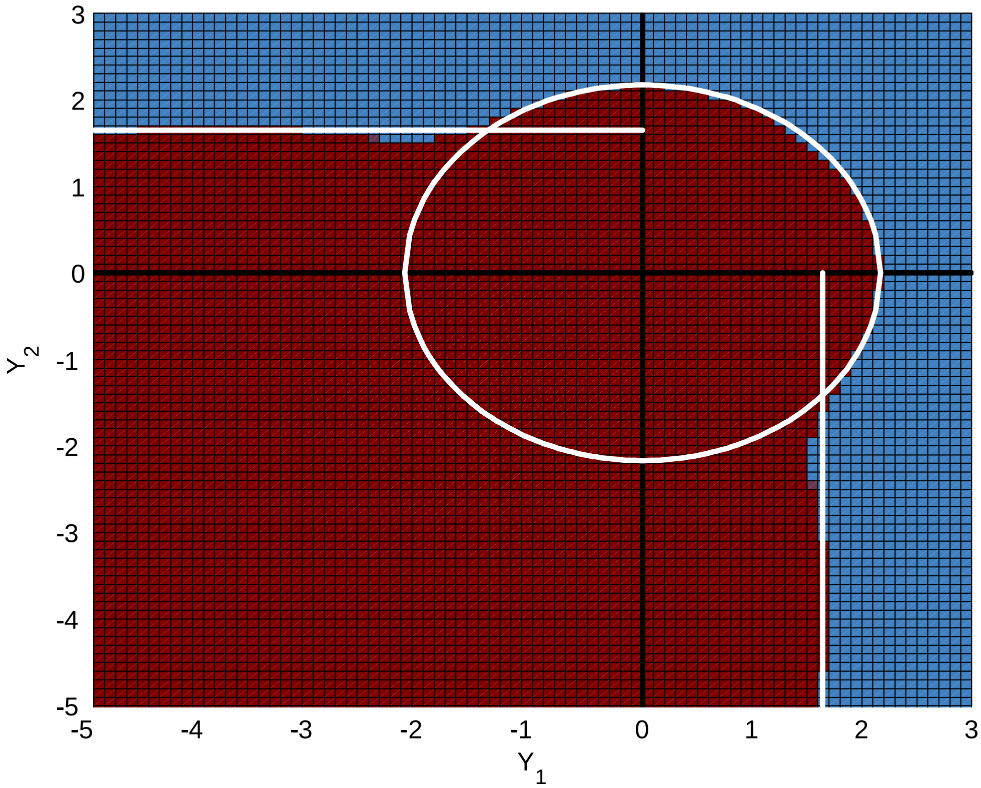

Figure 1: The red (dark) region is a discrete approximation of the acceptance region for the optimal test of $H_0 : \mu_1, \mu_2 \leq 0$ against a very flat prior on the alternative. This is well approximated by the area under the white lines and circle.

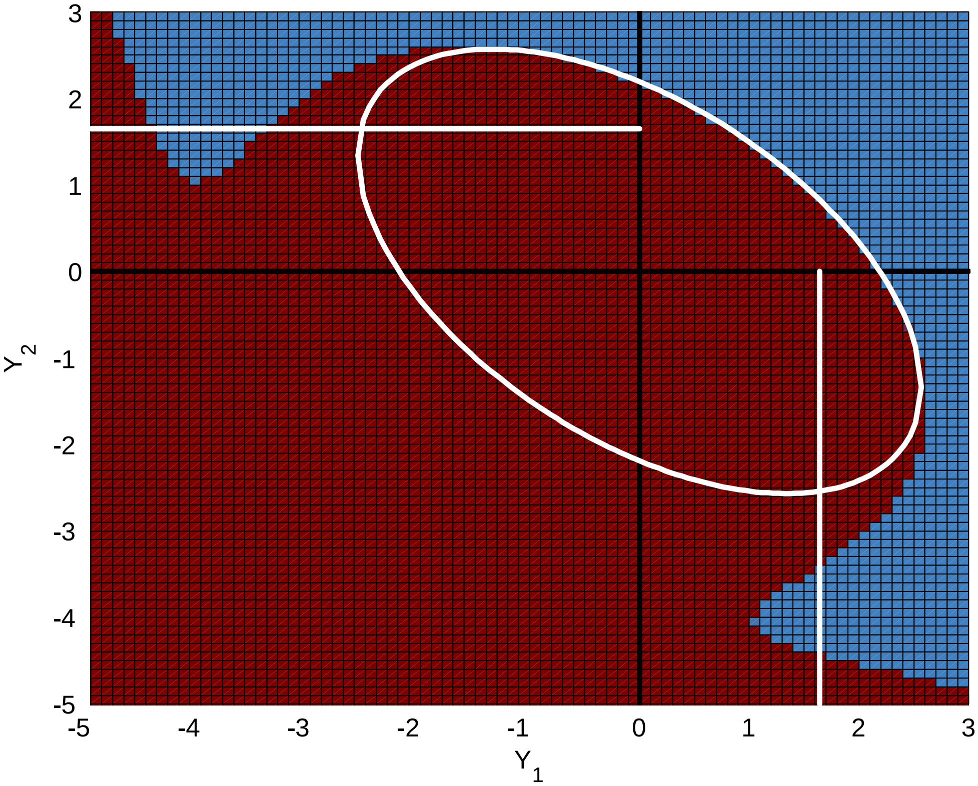

Figure 2: The red (dark) region is a discrete approximation of the acceptance region for the optimal test of $H_0 : \mu_1, \mu_2 \leq 0$ against a Gaussian alternative (13) with high correlation coefficient (V given by (14)). The area under the white ellipse and lines is the acceptance region of the test of Algorithm 3.1. Although there is disagreement in the two tests far from the origin, this makes little difference in terms of power since the weight under the alternative there is small.

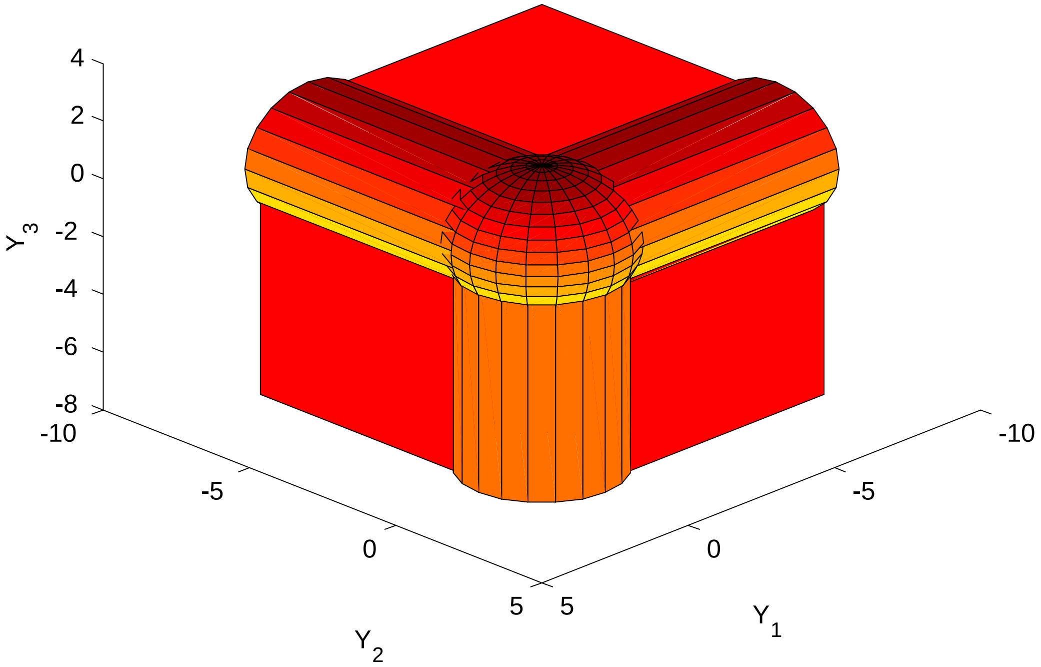

Figure 3: Acceptance region for Algorithms 3.1 and 3.2, for $\Sigma = I_3$ and flat distribution on the alternative.

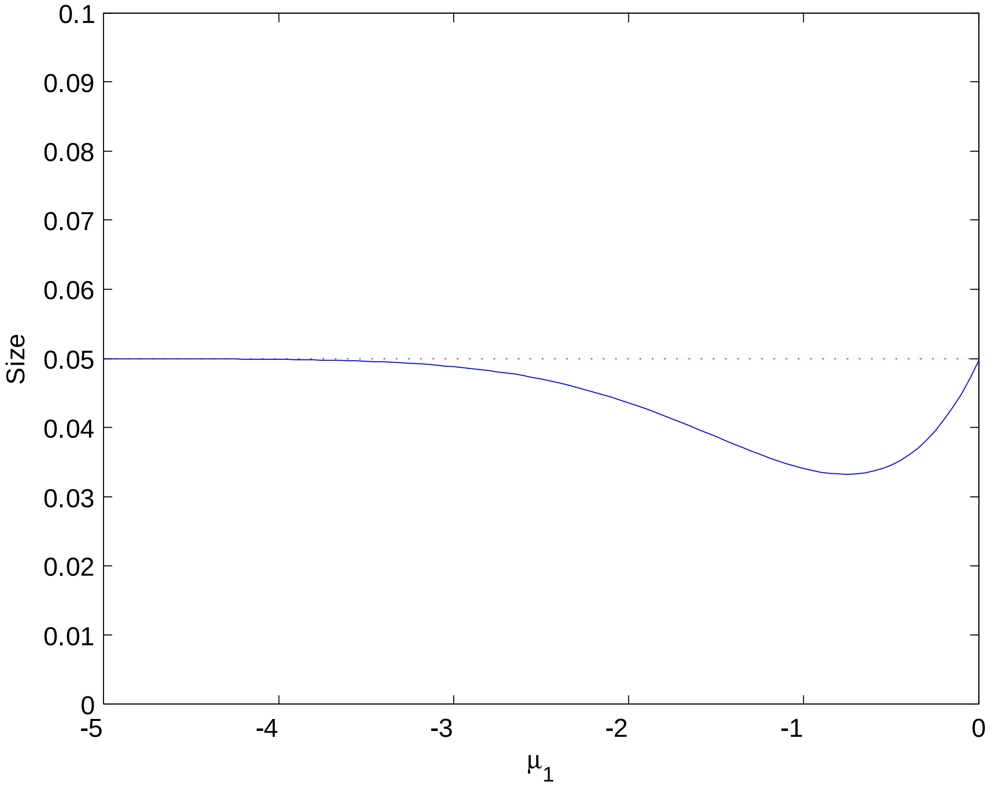

Figure 4: Size of test generated by Algorithms 3.1 and 3.2, for $\Sigma = I_2$, $V = I_2$, and nominal size $\alpha = 0.05$, at $\mu_2 = 0$ and various values of μ_1.

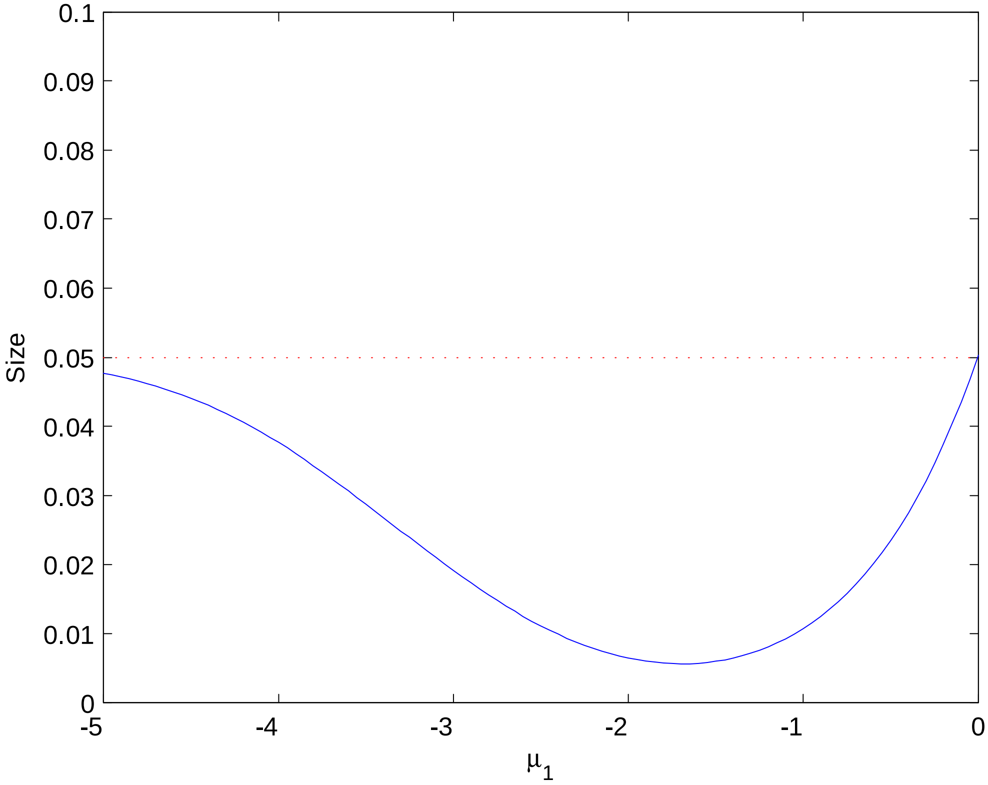

Figure 5: Size of test generated by Algorithms 3.1 and 3.2, for $\Sigma = I_2$, $V = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$, and nominal size $\alpha = 0.05$, at $\mu_2 = 0$ and various values of μ_1.

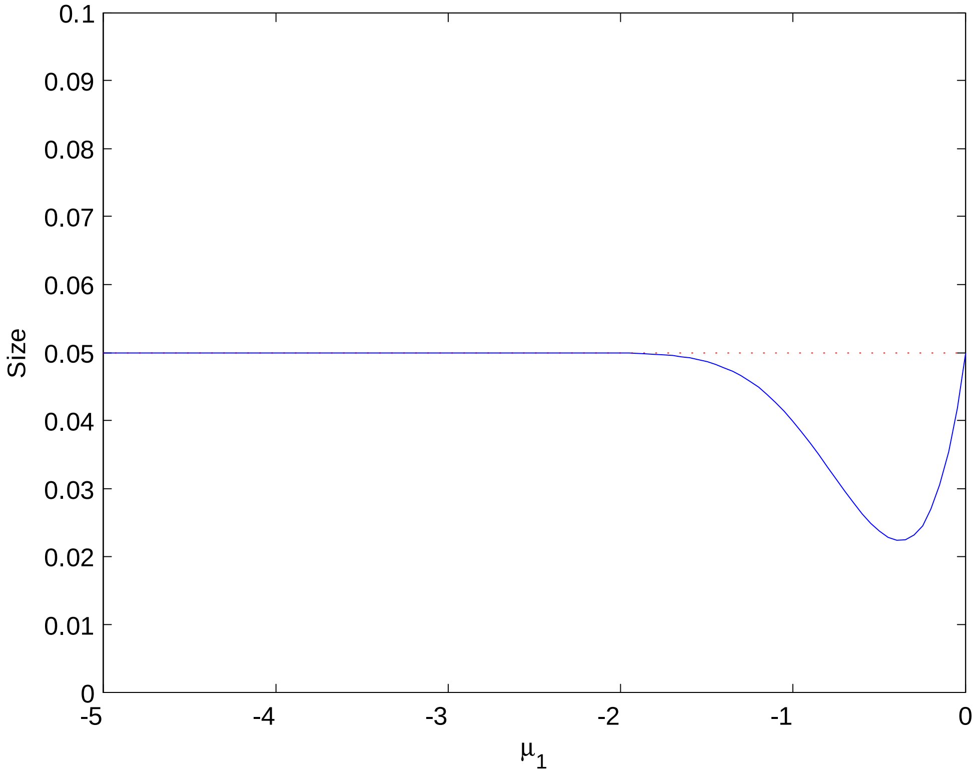

Figure 6: Size of test generated by Algorithms 3.1 and 3.2, for $\Sigma = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}$, $V = I_2$, and nominal size $\alpha = 0.05$, at $\mu_2 = 0$ and various values of μ_1.

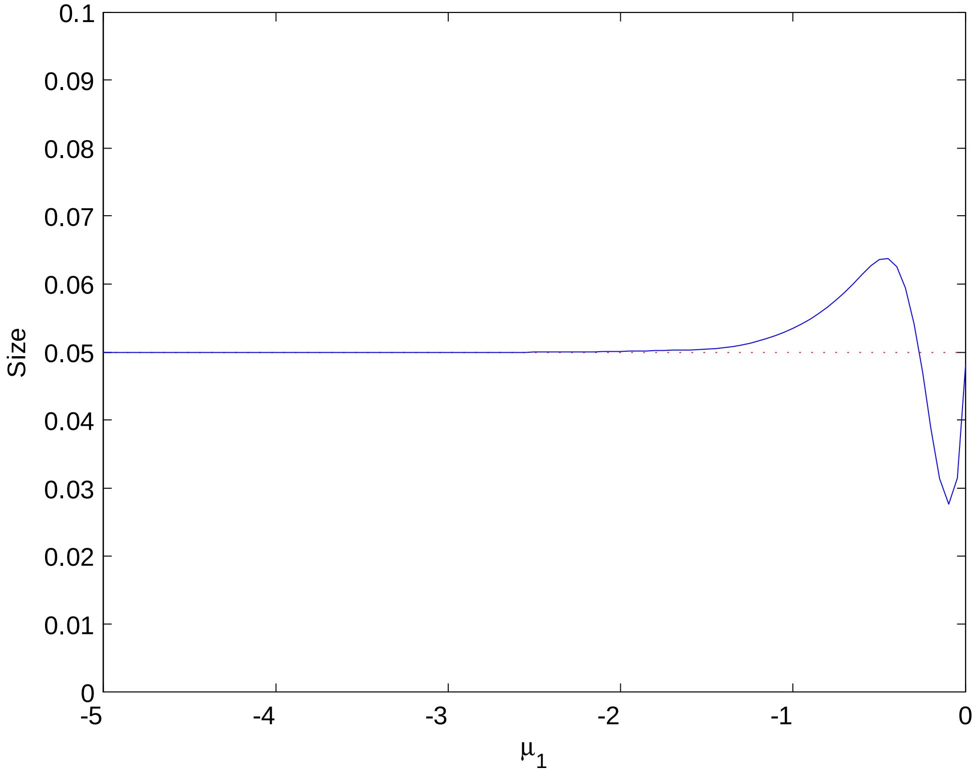

Figure 7: Size of test generated by Algorithms 3.1 and 3.2, for $\Sigma = \begin{pmatrix} 1 & -0.99 \\ -0.99 & 1 \end{pmatrix}$, $V = I_2$, and nominal size $\alpha = 0.05$, at $\mu_2 = 0$ and various values of μ_1.

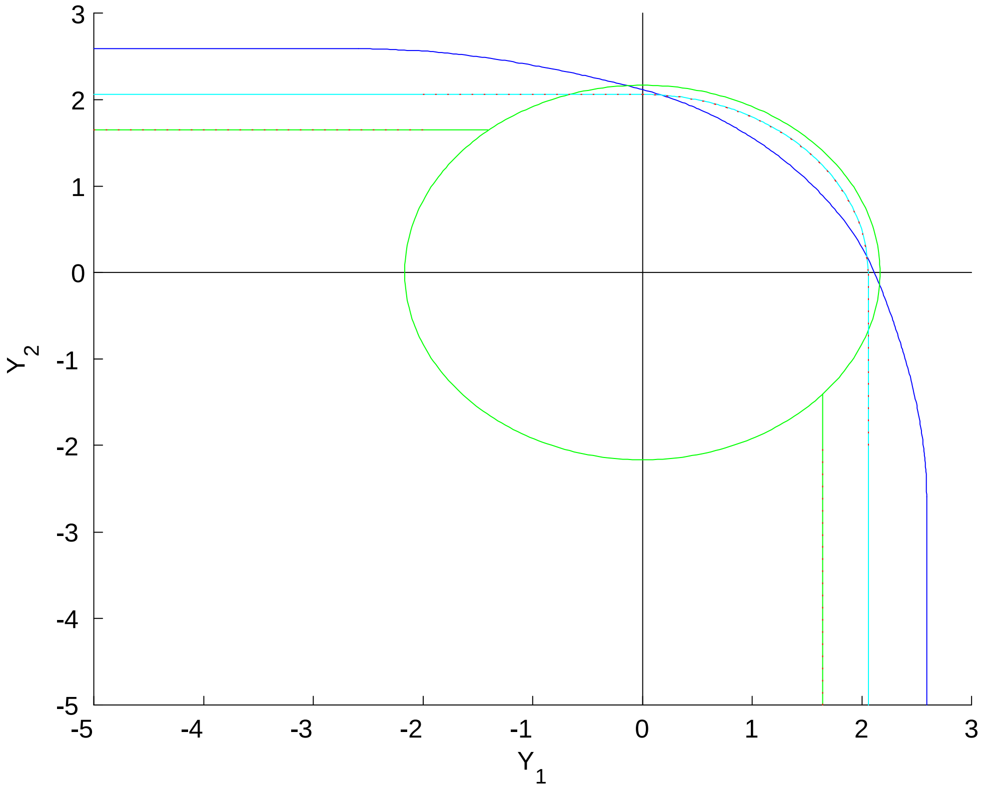

Figure 8: Acceptance regions for Rosen's (2007) test (cyan), Kim's (2008) test (blue) using the worst-case critical value, the Andrews and Soares's (2007) generalized moment selection adaptation of Rosen's test (red dotted lines), and the approximation to the optimal test against a flat prior shown in Figure 1 (green).