ADMISSIBLE, SIMILAR TESTS: A CHARACTERIZATION

José Luis Montiel Olea*

This paper studies a classical problem in statistical decision theory: a hypothesis test of a sharp null in the presence of a nuisance parameter. The main contribution of this paper is a characterization of two finite-sample properties often deemed reasonable in this environment: *admissibility* and *similarity*. Admissibility means that a test cannot be improved uniformly over the parameter space. Similarity requires the null rejection probability to be unaffected by the nuisance parameter.

The characterization result has two parts. The first part—established by Chernozhukov, Hansen, and Jansson (2009)—states that maximizing Weighted average power (WAP) subject to a similarity constraint suffices to generate admissible, similar tests. The second part—hereby established—states that constrained WAP maximization is (*essentially*) a necessary condition for a test to be admissible and similar. The characterization result shows that choosing an admissible, similar test is tantamount to selecting a particular weight function to report weighted average power. This result applies to full vector inference with a nuisance parameter, not to subvector inference.

The paper also revisits the theory of testing in the instrumental variables model. Specifically—and in light of the relevance of the weighted average power criterion in the main theoretical result—the paper suggests a weight function for the structural parameters of the homoskedastic instrumental variables model, based on the priors proposed by Chamberlain (2007). The corresponding test is, by construction, admissible and similar. In addition, the test is shown to have finite- and large-sample properties comparable to those of the conditional likelihood ratio test.

Keywords: Admissibility, hypothesis testing, instrumental variables, statistical decision theory, weak instruments.

*A previous version of this paper was circulated under the title "Efficient Conditionally Similar-on-the-Boundary Tests". I am deeply indebted to my main advisers Matías Cattaneo, Gary Chamberlain, Tomasz Strzalecki, and James Stock, for their continuous guidance, support, patience, and encouragement. I would like to thank seminar audiences at Banco de México, Brown, Chicago Booth (Econometrics), Cowles Summer Conference, Davis ARE, Duke, Imperial College Business School, ITAM, Northwestern, University of Chicago, University of Michigan, NYU Economics, Penn State, and TSE. I owe special thanks to Isaiah Andrews, Michael Jansson, Frank Schorfheide, Anna Mikusheva, Peter Phillips, Elie Tamer, Quang Vuong, and four anonymous referees for extremely helpful comments and suggestions. I would also like to thank Luigi Caloi, Hamza Husain, and Amilcar Velez for excellent research assistance. All errors are my own. First draft: September 18th, 2013. This version: May 14th, 2019.

1. INTRODUCTION

This paper studies hypothesis testing problems with a sharp null and a nuisance parameter. The main assumption is the existence of a parametric statistical model separately continuous in the parameters and the data. A common example of this environment is a testing problem concerning the coefficient of a right-hand endogenous regressor in an instrumental variables (IV) model.

This paper focuses on tests that satisfy two classical finite-sample properties: admissibility and similarity. *Admissibility* (Wald (1950)) is a weak optimality requirement for a test ϕ: there is no other test ϕ' with better rates of Type I and Type II error. *Similarity* (Neyman (1935)) is a stringent size-control condition for a test ϕ: the rate of Type I error of an α-similar test equals α regardless of the value of nuisance parameters.

The main result herein shows that admissible, similar tests are—*essentially*—weighted average power (WAP) maximizers inside the class of α-similar tests.[1] Thus, if a researcher finds admissibility and similarity attractive, the problem of selecting an admissible similar test is tantamount to selecting a particular weight function. This result applies to full vector inference with a nuisance parameter, not to sub-vector inference.

More formally, the characterization result has two parts. The first part states that maximizing WAP subject to a similarity constraint suffices to generate tests that are both admissible and similar (a property that was established in the work of Chernozhukov et al. (2009) in the context of the instrumental variables model). The second part—the main contribution of this paper—states that every admissible, similar procedure is an *extended* WAP-similar test (a concept that will be made precise). As a note for the theorists, the only assumption needed for the main result is a mild continuity requirement for the statistical model under consideration. The proof is based on an *essentially Complete Class Theorem* (see Theorem 2.9.2 and 2.10.3 in Ferguson (1967) and also Le Cam (1986), Chapter 2, Theorem 1). A key step is to show that the set of α-similar tests is compact in the weak* topology.

This paper also revisits the problem of conducting tests about the coefficient of a right-hand endogenous regressor in an instrumental variables model. In light of the

[1]WAP-similar tests have been used before in the context of the instrumental variables model to construct a power envelope for the conditional likelihood ratio test [Andrews, Moreira, and Stock (2006)] and to show that there is no conventional sense in which the Anderson and Rubin (1949) test wastes power in a linear homoskedastic instrumental variables model.

relevance of the weighted average power criterion in the characterization result, the paper studies the question of how to select a weight function in an overidentified homoskedastic model. Since there is no uniformly most powerful test in this environment, a weight function has practical relevance: it facilitates power comparisons among different similar tests via the weighted average power criterion, even if there is no final interest in implementing a WAP-similar test.

To illustrate this point, and building on the analysis of Chamberlain (2007), the paper presents a weight function, w^*, for the structural parameters of the homoskedastic IV model and derives a closed-form expression for the corresponding similar test that maximizes weighted average power (see Result 1). Analytical and numerical evidence suggests that the performance of the new test for the homoskedastic model is comparable to that of the conditional likelihood ratio (CLR) test of Moreira (2003), both asymptotically and in finite samples.[2] In addition (and by construction) the test is admissible and has larger weighted average power under w^* than the CLR, the Anderson and Rubin (1949) (AR) test, and the Lagrange multiplier (LM) test.

The remainder of this paper is organized as follows. Section 2 presents the finite-sample characterization of admissibility and similarity. Section 3 presents the weight function for the homoskedastic instrumental variables model. The main text Appendix states the lemmas used to establish the main results. Proofs, as required, are provided online in the supplementary material associated with this article, available at Cambridge Journals Online (journals.cambridge.org/ect)

2. FINITE-SAMPLE THEORY

The main definitions in this section follow Chamberlain (2007); Chapters 2 and 5 in Ferguson (1967); and Chapter 4 in Linnik (1968). The main result of this section is Theorem 1, which presents the characterization of admissibility and similarity.

2.1. *Basic Elements of a Testing Problem and Main Assumption*

The finite-sample parametric testing problem studied in this paper has the following components. There is a random vector X that takes values in the *sample space* $\mathbf{X} \subseteq \mathbb{R}^s$. There is a *parameter space* $\boldsymbol{\Theta} = B \times \boldsymbol{\Pi} \subseteq \mathbb{R}^{d_\beta \times d_\Pi} = \mathbb{R}^p$ whose ele-

[2]In finite samples, the test is invariant to rotation of the instruments, similar, and locally unbiased. In large samples, the test is equivalent to the CLR under strong-instrument asymptotics and valid under weak-instrument asymptotics.

ments $\theta = (\beta, \Pi) \in \mathbf{\Theta}$ are used to index a set of probability density functions (w.r.t. to Lebesgue measure in $\mathbb{R}^s$) over the sample space, $X \sim f(x;\theta)$. The collection $\{ f(x;\beta,\Pi) \}_{(\beta,\Pi)\in\mathbf{\Theta}}$ is called a *statistical model*.

The null hypothesis $\mathbf{H}_0$ states $X \sim f(x;\beta_0,\Pi)$ for some $\Pi \in \mathbf{\Pi}$: Π is a *nuisance parameter*. The null set Θ_0 is the set of parameters (β,Π) that satisfy $\mathbf{H}_0$. The alternative hypothesis $\mathbf{H}_1$ states $X \sim f(x;\beta,\Pi)$ for $\beta \neq \beta_0$, $\Pi \in \mathbf{\Pi}$. The alternative set Θ_1 is defined as $\mathbf{\Theta}\backslash\Theta_0$.[3] The testing problem studied in this paper is abbreviated as:

$$\mathbf{H}_0 : \beta = \beta_0 \text{ vs. } \mathbf{H}_1 : \beta \neq \beta_0.$$

A *test* is a measurable mapping $\boldsymbol{\phi} : \mathbf{X} \to [0,1]$. The scalar $\boldsymbol{\phi}(x)$ is interpreted as the probability of rejecting $\mathbf{H}_0$ after a realization x of X. The collection $\mathcal{C}$ will denote the class of all tests.

In a finite-sample set-up, tests are usually compared on the basis of the Type I and Type II error functions. The *rate of Type I error* of test $\boldsymbol{\phi}$ at $\theta \in \Theta_0$ is defined as $\mathbb{E}_\theta[\boldsymbol{\phi}(X)]$. The *rate of Type II error* of ϕ at $\theta \in \Theta_1$ is defined as $1 - \mathbb{E}_\theta[\boldsymbol{\phi}(X)]$. The rates of Type I and Type II error are typically summarized by the *risk function*:

$$R(\boldsymbol{\phi},\theta) \equiv \begin{cases} \mathbb{E}_\theta[\boldsymbol{\phi}(X)] & \text{if} \quad \theta \in \Theta_0 \\ \\ 1 - \mathbb{E}_\theta[\boldsymbol{\phi}(X)] & \text{if} \quad \theta \in \Theta_1. \end{cases}$$

Two tests $\boldsymbol{\phi}, \boldsymbol{\phi'}$ are said to be *risk equivalent* if $R(\boldsymbol{\phi},\theta) = R(\boldsymbol{\phi}',\theta)$ for all $\theta \in \Theta$.[4]

Assumption F0 below restricts the class of statistical models under consideration.

Assumption F0 (Separate Continuity): The statistical model $f(x;\theta)$ is:

i) continuous in θ for almost every $x \in \mathbf{X}$,

ii) continuous in x for almost every $\theta \in \Theta$.

[3] The null set is assumed to be nonempty, closed relative to the subspace topology in $(\Theta, \mathcal{T})$, and with an empty interior. This guarantees that Θ_0 coincides with its topological boundary $\text{Bd}\Theta_0$.

[4] This risk function implicitly assumes a "0-1" loss function, which is standard for testing problems; see Ferguson (1967), equation 5.4, p. 199.

2.2. *Admissibility, Similarity, and WAP-similar Tests*

This section presents the formal definitions of admissibility and similarity, along with a characterization of these properties in two-sided hypothesis problems in which the statistical model satisfies Assumption F0.

Admissibility: (Ferguson (1967), p. 54) The test $\boldsymbol{\phi}$ is *admissible* within the class $\mathcal{C}^* \subseteq \mathcal{C}$ if there is no $\phi' \in \mathcal{C}^*$ such that $R(\boldsymbol{\phi}', \theta) \leq R(\boldsymbol{\phi}, \theta)$ for all $\theta \in \boldsymbol{\Theta}$, with strict inequality for at least one $\theta \in \boldsymbol{\Theta}$.[5]

Admissibility was first introduced by Wald (1950) and it is a well-known concept in mathematical statistics. *This paper focuses on admissibility with respect to the class of all tests.*[6]

Similarity on Θ_0: A test $\boldsymbol{\phi}$ is α-*similar* on Θ_0 (α-s) if:

$$\mathbb{E}_\theta[\boldsymbol{\phi}(X)] = \alpha, \quad \forall\, \theta \in \Theta_0.$$

Similarity was first introduced by Neyman (1935) and it has been extensively studied by Linnik (1968). In two-sided problems with a nuisance parameter similarity provides a stringent size-control condition ('$=\alpha$' as opposed to '$\leq \alpha$').[7]

The main contribution of this paper is a characterization of admissibility and similarity. As mentioned in the introduction, part i) of the characterization result shows that WAP-similar tests are admissible in the class of all tests. Part ii) shows that every admissible test that is α-similar is *essentially* a WAP-similar test of level α. In order to state the main result and to formalize the notion of 'essentially' this paper

[5]Define an 'ordering' over tests as a binary relation $\succ$ in the space of all tests that verifies two properties. The first one is asymmetry: $\boldsymbol{\phi} \succ \boldsymbol{\phi}' \implies \boldsymbol{\phi}' \not\succ \boldsymbol{\phi}$. The second one is transitivity: $\boldsymbol{\phi} \succ \boldsymbol{\phi}'$ and $\boldsymbol{\phi}' \succ \boldsymbol{\phi}''$ implies $\boldsymbol{\phi} \succ \boldsymbol{\phi}''$. Admissibility induces an ordering through the "weakly dominated" binary relation: a test $\boldsymbol{\phi}'$ *weakly dominates* $\boldsymbol{\phi}$ if $R(\boldsymbol{\phi}', \theta) \leq R(\boldsymbol{\phi}, \theta)$ with strict inequality for at least one $\theta \in \boldsymbol{\Theta}$.

[6]Consequently, the notion of admissibility of a test herein discussed coincides with d-admissibility as defined in p. 233 of Lehmann and Romano (2005). See also problem 6.32 in the same book for details about the equivalence between admissibility and d-admissibility in testing problems.

[7]Similar tests always exist. The randomized test that rejects with probability α regardless of the data is, of course, α-similar but unfortunately has 'trivial' power. The main result in this paper suggests that WAP-similar tests are a reasonable class to search for similar tests with 'nontrivial' power. If the WAP-similar test happens to be of the form $\phi(x) = \alpha$, then our results imply that such a test cannot be improved uniformly over the parameter space.

introduces the following definitions. Let $w(\beta,\Pi)$ be a probability measure over $B\times\mathbf{\Pi}$.

WAP-similar tests: The α-similar test $\boldsymbol{\phi}_{\text{WAP}}^{w,\alpha}$ is (w)-WAP-similar if:

$$\text{WAP}(\boldsymbol{\phi}_{\text{WAP}}^{w,\alpha},w) \equiv \int_{\Theta_1}\Big(\int_{\mathbf{X}} \boldsymbol{\phi}_{\text{WAP}}^{w,\alpha}(x)f(x;\theta)dx\Big)dw(\theta) \geq \text{WAP}(\boldsymbol{\phi},w).$$

for any other α-similar test $\boldsymbol{\phi}$.

WAP-similar tests are indexed by a user-specified weight function $w(\beta,\Pi)$.[8] The weight function $w(\beta,\Pi)$ represents the part of the parameter space for which a WAP-similar test directs its power. This is particularly relevant in problems that do not admit a Uniformly Most Powerful (UMP) test.

The definition of an extended WAP-similar test is based on the classical notion of an extended Bayes test, as defined in Ferguson (1967) p. 50, Definition 3:

Extended WAP-similar Tests: The α-similar test $\boldsymbol{\phi}_{\text{E-WAP}}^{\alpha}$ is extended WAP-similar of level α if $\forall\, \epsilon > 0$ there exists a Borel probability measure $w_\epsilon(\beta,\Pi)$ supported on a non-empty subset of $\Theta_1 \equiv \Theta\backslash\Theta_0$ such that:

$$\text{WAP}(\boldsymbol{\phi}_{\text{WAP}}^{w_\epsilon,\alpha},w_\epsilon) \geq \text{WAP}(\boldsymbol{\phi}_{\text{E-WAP}}^{\alpha},w_\epsilon) \geq \text{WAP}(\boldsymbol{\phi}_{\text{WAP}}^{w_\epsilon,\alpha},w_\epsilon) - \epsilon.$$

Extended WAP-similar tests are *essentially* WAP-similar tests: for any $\epsilon > 0$ there is a weight function $w_\epsilon(\beta,\Pi)$ such that the $w_\epsilon(\beta,\Pi)$-WAP of any other α-similar test can exceed that of the extented test by at most ϵ. So, up to the tolerance ϵ, extended WAP-similar tests essentially maximize weighted average power.

Theorem 1: *Suppose Assumption F0 holds.*

i) If $w(\beta,\Pi)$ has a p.d.f. with full-support on Θ_1, $\boldsymbol{\phi}_{\text{WAP}}^{w,\alpha}$ is admissible and α-similar.

ii) Every admissible, α-similar test is an extended WAP-similar test of level α.

In both i) and ii), admissibility is taken with respect to the class of all tests.

[8] For notational convenience, this paper uses $\boldsymbol{\phi}_{\text{WAP}}(x_1,x_2)$ instead of $\boldsymbol{\phi}_{\text{WAP}}^{w,\alpha}(x_1,x_2)$ whenever convenient.

PROOF: See Appendix A.1 *Q.E.D.*

COMMENT ON PART i): The first part of Theorem 1 says that a test that maximizes power *only* among α-similar tests, is in fact admissible with respect to *all* tests. The proof of part i) is straightforward and generalizes the results in Chernozhukov et al. (2009). First, ignoring 'oversized' tests is inconsequential: tests with rate of Type I error larger than α can never dominate an α-similar test. Second, there is no α-similar test that dominates $\phi^{w,\alpha}_{\text{WAP}}$ (this follows from the WAP maximization property, provided the weight function has full-support). Finally, Assumption F0 is used to show that the continuity of the risk function implies that a constrained WAP maximizer cannot be dominated by a nonsimilar procedure of size α.

COMMENT ON PART ii): To the best of my knowledge part ii) of Theorem 1 is new. The proof of ii) is based on an *essentially complete class theorem* [see Theorem 2.9.2 and 2.10.3 in Ferguson (1967) and also Le Cam (1986), Chapter 2, Theorem 1]. Broadly speaking, the theorem states that if C is an *essentially complete class* relative to D (with $C \subseteq D$) and C is compact (relative to some topology that makes the risk function continuous), then *Extended Bayes rules* in C are essentially complete relative to D. Essential completeness of a class C (relative to D) means that for any rule in D, there is a rule in C with smaller than or equal risk; see Ferguson (1967), p. 55, Definition 3. The definition of an extended Bayes rule is presented in Ferguson (1967) p. 50, Definition 3.

A key lemma in the proof of the paper shows that the set of α-similar tests—denoted $C(\alpha\text{-s})$—is compact in the weak* topology (see Lemma 1 in Appendix A.1). Since the risk function of the testing problem (Type I and Type II error) is continuous with respect to the same topology, the essentially complete class theorem applies: this is, the set of extended Bayes tests in $C(\alpha\text{-s})$ is an essentially complete class relative to $C(\alpha\text{-s})$ (which is essentially complete relative to itself). The essential completeness of extended Bayes tests in $C(\alpha\text{-s})$ implies that any admissible test in $C(\alpha\text{-s})$ must be an extended Bayes test. Extended Bayes tests in $C(\alpha\text{-s})$ are then shown to be extended WAP-similar tests. The desired result follows.

REMARK 1: Finding a WAP test typically requires a researcher to solve an optimization problem over an infinite dimensional choice set. It is well-known that the α-level (w)-WAP-similar Test for a two-sided testing problem where the data can be partitioned as $X = (x_1, x_2)$ and x_2 is a *boundedly-complete, null-sufficient statistic* (i.e., $f(x_1|x_2;\beta_0) \equiv f(x_1|x_2;\beta_0,\Pi) = f(x_1|x_2;\beta_0,\Pi')$ for all Π) is given by:

(2.1)

$$\phi^{w,\alpha}_{\text{WAP}}(x_1,x_2) = \begin{cases} 1 & \text{if} \quad z(x_1,x_2) \equiv f^*_w(x_1,x_2)/f(x_1|x_2;\beta_0) > c(x_2;\alpha) \\ 0 & \text{otherwise} \end{cases},$$

where

$$f^*_w(x_1,x_2) \equiv \int_{B\times\mathbf{\Pi}} f(x_1,x_2;\beta,\Pi)dw(\beta,\Pi) < \infty \quad \forall\,(x_1,x_2), \tag{2.2}$$

and for each x_2, $c(x_2;\alpha)$ corresponds to the conditional (1-α) quantile of the random variable $z(X_1,x_2)$, with $X_1 \sim f(x_1|x_2;\beta_0)$. See Lemma 5 in Appendix B.3.2 for the definition of a bounded-complete, null-sufficient statistic and the derivation of equation (2.1).

Remark 2: Under the additional assumption that $\Theta \subseteq \mathbb{R}^p$ is compact, it is possible to present a slightly stronger version of Theorem 1 part ii). Let $\to^*$ denote convergence in the weak* topology as defined in Section B.1.1 of the Online Supplementary Material. One can show that for every admissible, α-similar test ϕ there exists a sequence of Borel probability measures w_n on Θ and an α-similar test ϕ^* such that $w_n \xrightarrow{d} w^*$, and $\phi^{w_n,\alpha}_{\text{WAP}} \to^* \phi^*$. Moreover, if the probability measures w_n admit some common dominating measure and their Radon-Nikodym derivatives are bounded by a common integrable function, then it is possible to show that the WAP of ϕ and ϕ^* coincide under w^*. Thus, it is possible to claim that—if some additional regularity conditions are met—every admissible, similar test ϕ can be shown to be 'WAP equivalent' to a properly defined limit of WAP similar tests. This result is formalized as a corollary to Theorem 1 in Appendix B.3.1.

3. INSTRUMENTAL VARIABLES MODEL

In light of the relevance of the WAP criterion in the characterization result, this section revisits the question of how to construct a weight function for the structural parameters of a homoskedastic, over-identified linear IV model.

Of course, the choice of weights is idiosyncratic. In fact—in light of Theorem 1—there is no choice that could be deemed better than another, if admissibility and similarity are the only properties of interest to the researcher. The objective of this section is to provide context-free weights for a homoskedastic IV model that translate into a test with desirable finite- and large- sample properties. The weights can also be used for WAP comparisons and for understanding how much average power is compromised by using conventional testing procedures as opposed to WAP maximizing tests.

The outline of this section is as follows. First, we present the weights w^* for the structural parameters of a homoskedastic IV model. Second, we derive the WAP-similar test associated to w^* and study some of its finite– and large–sample properties. One of the main findings is that the WAP-similar test is 'close' to the conditional likelihood ratio (CLR) of Moreira (2003). Finally, in Section B.4.1 of the Appendix, we derive the large-sample properties of the new test.

Set-up and Notation: In a sample of size n, let $y \in \mathbb{R}^n$ denote a vector that collects the outcome variable for each observation in the sample. Likewise, $x \in \mathbb{R}^n$ denotes the vector that collects all the observations of the endogenous regressor and $Z \in \mathbb{R}^{n\times k}$ is the matrix that collects the k instrumental variables. The parameter β denotes the coefficient of the endogenous regressor of interest and Π denotes the first-stage coefficient. The testing problem is:

$$\mathbf{H}_0 : \beta = \beta_0 \quad \text{vs.} \quad \mathbf{H}_1 : \beta \neq \beta_0,$$

where β_0 is some prespecified value and Π, the first-stage coefficient, is the nuisance parameter.

Statistical Model for IV: Let $\widehat{\gamma}_n$ denote the OLS estimator of the reduced-form parameters in the IV regression. The parametric statistical model specifies a multivariate normal distribution for $\widehat{\gamma}_n$:

$$\widehat{\gamma}_n \sim \mathcal{N}_{2k}\left(\begin{pmatrix} \beta\Pi \\ \Pi \end{pmatrix}, \frac{\Sigma}{n} \right), \quad \text{where } \widehat{\gamma}_n \equiv \begin{pmatrix} (Z'Z)^{-1}Z'y \\ (Z'Z)^{-1}Z'x \end{pmatrix}. \tag{3.1}$$

Thus, the application of the theory in the first part of the paper will require the Gaussian model above to approximate the finite-sample distribution of $\widehat{\gamma}_n$.[9] Fortunately, a central limit theorem (CLT) and a local-to-zero first-stage as in Staiger and Stock (1997) guarantees that (3.1) will be, indeed, a reasonable approximation for a wide range of data generating processes (including stochastic instruments, time varying heteroskedasticity, autocorrelation of unknown form, etc).[10]

In models with i.i.d. conditionally homoskedastic data, the covariance matrix Σ can be written as $\Omega \otimes Q^{-1}$, where Ω is the matrix of second moments of reduced-form residuals, Q is the matrix of second moments of the instrumental variables, and $\otimes$ denotes the Kronecker product. This 'Kronecker' case is commonly used to analyze the power properties of tests in the IV model, and it is the focus of this section.

Weights for (β, Π) in the Homoskedastic IV-Model: Let ρ denote a random variable supported on $\mathbb{R}_+$. Let ϕ denote a random vector supported on $\mathcal{S}^1$, the unit sphere. Let ω denote a random vector supported on $\mathcal{S}^{k-1}$, the $k-1$ sphere.[11] Let $a_0 = (\beta_0, 1)'$, $b_0 = (1, -\beta_0)'$. If Σ is of the form $\Psi \otimes \Phi$, consider weights on (β, Π) of the form:

$$\begin{pmatrix} \beta\Pi \\ \Pi \end{pmatrix} = n^{-1/2} \left(\Psi C_0' \otimes \Phi^{1/2} \right) \rho(\phi \otimes \omega), \; C_0 \equiv \begin{pmatrix} (b_0'\Psi b_0)^{-1/2} b_0' \\ (a_0'\Psi^{-1}a_0)^{-1/2} a_0'\Psi^{-1} \end{pmatrix}, \tag{3.2}$$

where ϕ and ω are two independent random variables, uniformly distributed in their domain; this is:

$$\phi \sim \mathcal{U}(\mathcal{S}^1), \quad \omega \sim \mathcal{U}(\mathcal{S}^{k-1}). \tag{3.3}$$

---

[9]This parametric model has been considered before by Müller (2011), p. 423; Moreira and Moreira (2015), p. 10; and I.Andrews (2016), p. 6.

[10]The distribution theory in Staiger and Stock (1997) relies on an assumed CLT for the inner product of instruments and reduced-form residuals. The actual CLT required is given and proved in Phillips (1989) (Lemma 2.3 and Theorem 2.4).

[11]For any $m \in \mathbb{N}$, $\mathcal{S}^m$ is the m unit sphere; that is $\mathcal{S}^m = \{x \in \mathbb{R}^{m+1} | \, ||x|| = 1\}$.

Finally, set:

$$\rho \sim \sqrt{\chi^2_k}, \tag{3.4}$$

with χ^2_k independent of ϕ and ω. These weights are based on a convenient reparameterization of the IV model analyzed in Chamberlain (2007).[12] In such reparameterization ω and ϕ represent—respectively—the 'direction' of the first-stage coefficient, Π, and also the direction of the alternative hypothesis, $(\beta, 1)'$. The parameter ρ represents the product of the norms of each of these vectors and it can be thought of as a measure of instrument strength.

Chamberlain (2007) considered *uniform* priors for ϕ, ω (uninformative priors for the directions) but left the distribution of ρ unspecified. The priors for ϕ and ω were motivated by studying a minimax problem (they are least-favorable). These priors aimed to be context-free; consequently, researchers should not expect any "hidden economics" buried in the weight function. Other prior specifications could be available depending on the application.

Appendix B.4.3 shows that the weights for (β, Π) proposed in this section—which are Chamberlain (2007)'s weights with the additional assumption that $\rho \sim \sqrt{\chi^2_k}$—coincide with the MM2 weights proposed by Moreira and Moreira (2015) (up to a scaling parameter) for the case in which Σ has a Kronecker form.

Distribution of $\sqrt{\lambda}(\beta - \beta_0)$ and λ: The i.i.d. homoskedastic model, which has a covariance matrix Σ of the form $\Omega \otimes (Z'Z/n)^{-1}$, has been analyzed in detail in the work of Andrews et al. (2006). The Monte-Carlo exercises reported in that paper depend on the parameters:

$$\lambda \equiv n\Pi'\Phi^{-1}\Pi, \text{ and } \sqrt{\lambda}(\beta - \beta_0).$$

The probability density function of $(\sqrt{\lambda}(\beta - \beta_0), \lambda)$ is given in Figure 1 below.

[12] The original parameters (β, Π) induce the following "canonical" parameters (ρ, ϕ, ω):

$$\rho = (a'\Omega^{-1}a)^{1/2}(\beta'Z'Z\beta)^{1/2}, \ \phi = C_0 a/(a'\Omega^{-1}a)^{1/2}, \ \omega = (Z'Z)^{1/2}\beta/(\beta'Z'Z\beta)^{1/2}.$$

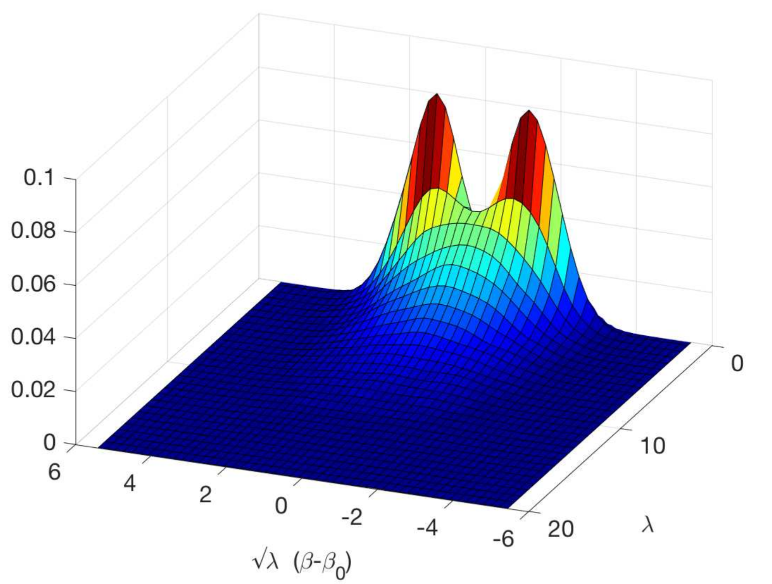
This yields the following parameterization of the IV model:

$$\begin{pmatrix} S_n \\ T_n \end{pmatrix} \sim \mathcal{N}_{2k}\left(\rho(\phi \otimes \omega), \quad \mathbb{I}_{2k} \right), \quad \rho \in \mathbb{R}_+, \quad \phi \in \mathcal{S}^1, \quad \omega \in \mathcal{S}^{k-1},$$

where S_n and T_n are defined in the statement of Result 1.

The WAP-similar test directs its power according to such density.

Figure 1: Weights for $(\sqrt{\lambda}(\beta - \beta_0), \lambda)$ induced by the weights on (ρ, ϕ, ω)



DESCRIPTION: Figure 1 is based on weights (3.3), (3.4) for the parameters (ϕ, ω, ρ) and the formulas in equations (B.35), (B.36) in the Appendix. The matrix Ψ is assumed to have unit diagonal elements and correlation parameter $r = .5$. The matrix Φ is assumed to be the identity. The null hypothesis is $\beta_0 = 0$. There are 4 instruments, $k = 4$. The bivariate density is generated by a Monte-Carlo exercise with 50,000 independent draws from (ρ^2, ϕ) and Matlab's bivariate density estimator `gkde2`. The script used to generate this figure is `Weights_WAP2016.m`.

The following result shows that, under the proposed weights, there is a closed-form solution for the test statistic of the corresponding WAP-similar test.

Result 1 (WAP-similar test for Homoskedastic IV): *Suppose that $\Sigma = \Psi \otimes \Phi$ where $\Psi \in \mathbb{R}^{2\times 2}$ and $\Phi \in \mathbb{R}^{k\times k}$ are positive definite, symmetric matrices. The α-WAP similar test for the problem $\boldsymbol{H}_0 : \beta = \beta_0$ vs. $\boldsymbol{H}_1 : \beta \neq \beta_0$ given the statistical model (3.1) and the weights over (β, Π) induced by (3.3), (3.4) rejects the null hypothesis if the statistic $z_{\text{WAP}}(S_n, T_n)$:*

$$(S_n'S_n - T_n'T_n) + 8\ln\Big(I_0\Big(\frac{1}{8}\Big[(S_n'S_n - T_n'T_n)^2 + 4(S_n'T_n)^2)\Big]^{1/2}\Big)\Big) + 4\ln(2\pi) + 4\ln((1/8)T_n'T_n)$$

exceeds the critical value function $c_{\text{WAP}}(T_n, \alpha)$. The critical value function is defined as the $(1-\alpha)$ quantile of the distribution of the statistic above with $S \sim \mathcal{N}_k(\boldsymbol{0}, \mathbb{I}_k)$ and T_n fixed. The function $I_0(\cdot)$ is the modified Bessel function of the first kind of order zero defined in Section 9.6, p. 374 of Abramowitz and Stegun (1964). The value $3.1415\ldots$ is denoted as π. The statistics S_n and T_n are given by:

$$\begin{pmatrix} S_n \\ T_n \end{pmatrix} \equiv$$

$$\begin{pmatrix} ([b_0' \otimes \mathbb{I}_k)\Sigma(b_0 \otimes \mathbb{I}_k)]^{-1/2} & \boldsymbol{0} \\ \boldsymbol{0} & [(a_0' \otimes \mathbb{I}_k)\Sigma^{-1}(a_0 \otimes \mathbb{I}_k)]^{-1/2} \end{pmatrix} \begin{pmatrix} (b_0' \otimes \mathbb{I}_k) \\ (a_0' \otimes \mathbb{I}_k)\Sigma^{-1} \end{pmatrix} \sqrt{n}\widehat{\gamma}_n,$$

with $a = (\beta, 1)'$, $a_0 = (\beta_0, 1)'$, $b_0 = (1, -\beta_0)'$ and Σ of the form $\Psi \otimes \Phi$.

Proof: See Appendix A.2. *Q.E.D.*

Comment on Result 1: If the expression:

$$8\ln\Big(\,I_0\Big(\frac{1}{8}\Big[(S_n'S_n - T_n'T_n)^2 + 4(S_n'T_n)^2)\Big]^{1/2}\Big)\,\Big)$$

were to be replaced by:

$$\Big((S_n'S_n - T_n'T_n)^2 + 4(S_n'T_n)^2)\Big)^{1/2},$$

the test in Result 1 would be equivalent to the conditional likelihood ratio test. Note that for large values of $T_n'T_n$ it is possible to motivate the desired substitution, up to an error term. In fact, it is common practice to evaluate the modified Bessel function $I_0(x)$ using the asymptotic approximation in Olver (1997), p. 435:

$$I_0(x) \Bigg/ \frac{e^x}{(2\pi x)^{1/2}} \to 1, \text{ as } x \to \infty,$$

which implies that for any s

$$z_{\text{WAP}}(s,t) - 2\text{CLR}(s,t) \to 0, \text{ as } t't \to \infty.$$

See Appendix B.4.4 for details. To further analyze the similarities (and differences) between the test in Result 1 and the CLR, we report a) the correlation between the two tests, b) conventional power plots, and the c) weighted average power of the two tests.

Correlation Between the WAP-similar test and the CLR: Under the null hypothesis, any nonrandomized test is a Bernoulli random variable with success probability equal to its rate of Type I error. Thus, one way to understand whether the CLR is ‘close’ to the test in Result 1 is by reporting the correlation between the two tests under the null hypothesis. Figure 4 suggests that the test in Result 1 and the CLR have a high correlation under the null hypothesis.

Power Plots: The analysis of correlation has nothing to say about the similarities (or differences) in power performance. Figure 2 presents a standard power comparison between the WAP-similar test in Result 1 and the Anderson and Rubin (1949) test (AR), the Lagrange multiplier Test (LM), the CLR, and the POSI2 (Point-Optimal Similar, Invariant Two-sided test). The figure reports the average power computed over the grid of values of $\lambda^{1/2}(\beta-\beta_0)$ using uniform weights and fixing the value of λ under consideration ($\lambda = n\Pi'\Phi^{-1}\Pi$). The figures also contain an (infeasible) power curve of the test that rejects whenever $|\Pi'\Phi^{-1/2}S|/(\Pi'\Phi^{-1}\Pi)^{1/2} > 1.96$. Figure 2 suggests that the power curves of the WAP-similar test in Result 1 and the CLR are almost indistinguishable. Also, the figure shows that there are alternatives (β, λ) for which the curve of the POSI2 test seems closer to the AR test than to the CLR.

Weighted Average Power Comparsion: The weights that define the WAP-similar test in Result 1 facilitate the numerical comparison of different testing procedures. For instance, in the context of Figure 2, the WAP comparison using weights (3.3) and (3.4) is as follows: 24% for the WAP-similar test, 23.8% for the CLR, 22.2% for the AR and 18.5% for the LM. Figure 3 compares the WAP of the WAP-similar test against that of the CLR, LM, and AR for a wider range of reduced-form correlations ($r \in [-.9, .9]$). One can claim that with 4 instruments the CLR is .15% away from maximizing weighted average power among the class of α-similar tests, given the proposed weights in (3.3) in (3.4). We also report the WAP using a uniform distribution over the area displayed in Figure 1. The WAP comparison is as follows: 62.33% for the WAP-similar test; 62.24% for the CLR; 56.10% for the LM; and 59.02% for the AR.

Conditional Rejection Region: The WAP-similar test in Result 1 (R1) is measurable with respect to the triplet $(\text{AR}_n, \text{LM}_n, T_n'T_n) \equiv (S_n'S_n, (S_n'T_n)^2/T_n'T_n, T_n'T_n)$. It is natural to ask whether the WAP-similar test rejects the null hypothesis when both the AR_n and LM_n do. Figure 5 reports 'conditional' critical regions in the (AR, LM) space for two different values of $T_n'T_n$. The conditional critical regions suggest that the WAP-similar test in Result 1 can be well approximated by a linear Combination between the AR_n and the LM_n as the tests in I. Andrews (2016).

Finite-Sample properties: The test in Result 1 is admissible and α-similar by construction. The test is invariant, for it depends on the data only through the triplet (AR_n, LM_n, $T_n'T_n$). Also, it is well known that any α-similar test that depends on the data only through such a triplet is locally unbiased; see Andrews et al. (2006), p. 730.

Large-Sample properties: The test in Result 1 was derived under the assumption that the rotated reduced-form OLS estimators $(S_n', T_n')'$ have the exact distribution:

$$Q^n_{\beta,\Pi,\Sigma} \equiv \mathcal{N}_{2k}\left(\begin{matrix} ([b_0' \otimes \mathbb{I}_k)\Sigma(b_0 \otimes \mathbb{I}_k)]^{-1/2}(\beta-\beta_0)\sqrt{n}\Pi \\ [(a_0' \otimes \mathbb{I}_k)\Sigma^{-1}(a_0 \otimes \mathbb{I}_k)]^{-1/2}(a_0' \otimes \mathbb{I}_k)\Sigma^{-1}(a \otimes \mathbb{I}_k)\sqrt{n}\Pi \end{matrix}, \; \mathbb{I}_{2k}\right),$$

where Σ is of the form $\Psi \otimes \Phi$. In any finite sample, however, the law of $(S_n', T_n')'$ is a function of (β, Π), the sample size, and the joint distribution between the instrumental variables and reduced-form residuals, denoted F. In fact, one can write:

$$\begin{pmatrix} ([b_0' \otimes \mathbb{I}_k)\widehat{\Sigma}(b_0 \otimes \mathbb{I}_k)]^{-1/2}(b_0' \otimes \mathbb{I}_k) \\ [(a_0' \otimes \mathbb{I}_k)\widehat{\Sigma}^{-1}(a_0 \otimes \mathbb{I}_k)]^{-1/2}(a_0' \otimes \mathbb{I}_k)\widehat{\Sigma}^{-1} \end{pmatrix} \sqrt{n}\widehat{\gamma}_n \sim P^n_{\beta,\Pi,F},$$

where $\widehat{\Sigma}$ is an estimator of the variance of $\sqrt{n}\widehat{\gamma}_n$. This variance depends on F and such dependence is denoted $\Sigma(F)$. The estimator $\widehat{\Sigma}$ need not have the Kronecker form, even when $\Sigma(F)$ does.

If one assumes that for n large enough the distributions P^n and Q^n are close to each other (under the null), then one would expect the rate of Type I error computed under P^n to be close to that obtained under Q^n. We formalize this statement in Appendix B.4.1. We also show that the test in Result 1 is as powerful (locally) as the GMM-Wald test for β_0 (Appendix B.4.2) based on the sample moment condition:

$$\frac{1}{\sqrt{n}} Z'(y - \beta_0 x) = \mathbf{0}.$$

Weights for the Heteroskedastic/Autocorrelated IV model: While it is possible to extend the weights w^* to an IV model with heteroskedasticity and autocorrelation (HAC), there is no guarantee such weights will lead to desirable finite- and large-sample power properties (such as unbiasedness and local power efficiency). In fact, I. Andrews (2016) and also Moreira and Moreira (2015) have used Yogo (2004)'s Elasticity of Intertemporal Substitution estimation exercise to analyze the finite-sample performance of different tests in the HAC-IV model. These papers find that WAP-similar tests can be badly biased and that the (quasi)-conditional likelihood ratio test can have poor performance in certain parts of the parameter space. Contrary to the homoskedastic case, the question of whether there exists an admissible, similar test with desirable finite- and large-sample power properties does not seem to have a straightforward answer. Theorem 1 implies, however, that if such a test exists it will have to be (essentially) a WAP-similar test.

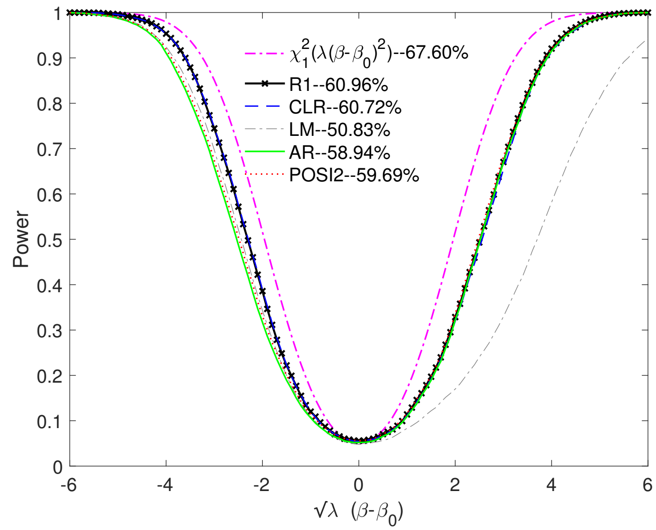
4. CONCLUSION

This paper studied testing problems with a sharp null and a nuisance parameter. The leading example was a testing problem concerning the coefficient of a single right-hand endogenous regressor in an instrumental variables (IV) model.

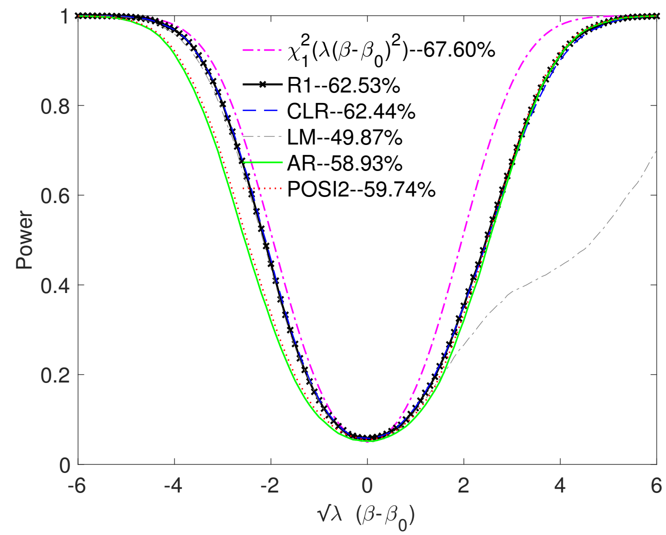
The main result in this paper showed that WAP-similar tests characterize two important finite-sample properties: *admissibility* and *similarity*. The characterization result stated that WAP-similar tests are admissible and similar; but more importantly, that every admissible, similar procedure is *essentially* a WAP-similar test. Thus, if a researcher finds admissibility attractive and at the same time desires to make the rate of Type I error invariant to nuisance parameters the WAP-similar class should be of interest. This result applies to full vector inference with a nuisance parameter, not to subvector inference.

In light of the relevance of the WAP criterion in the characterization result, this paper proposed a weight function w^* for the homoskedastic IV model. Result 1 showed that there is a closed-form expression for the test statistic of the WAP-similar test corresponding to w^*. Analytical and numerical evidence showed that the WAP-similar test has finite- and large- sample properties comparable to those of the conditional likelihood ratio of Moreira (2003). In addition to these properties, the WAP similar test is—by construction—admissible.

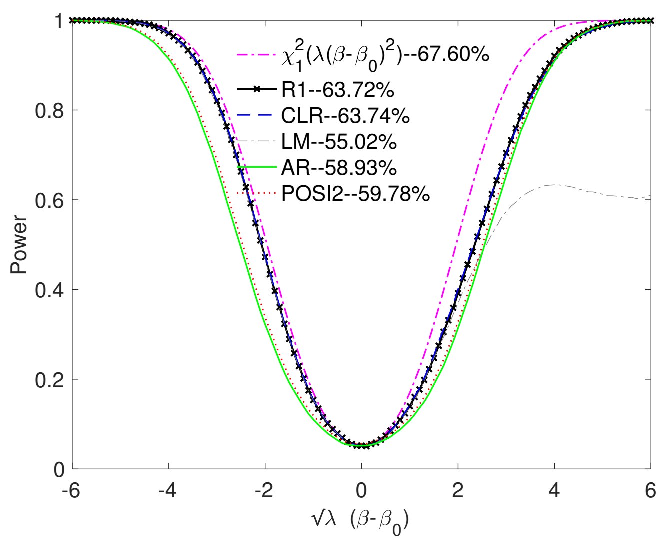
Figure 2: WAP-similar test (R1) vs. CLR, LM, AR, POSI2 (Power Plots)
($k = 4$, $r = .5$)



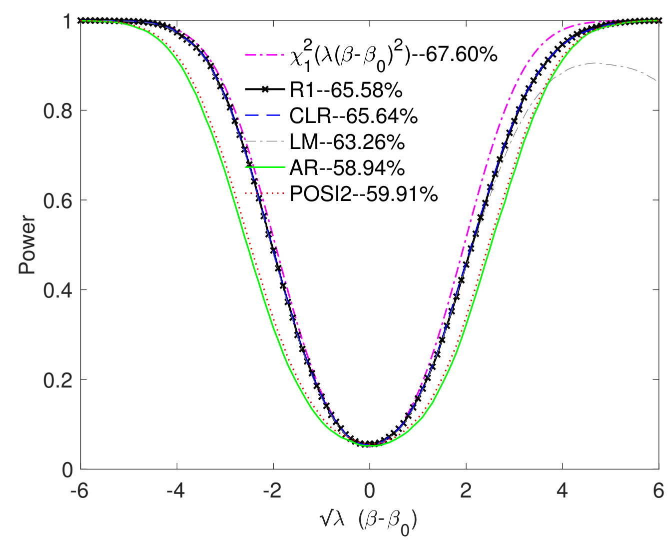
(a) $\lambda = 1$



(b) $\lambda = 5$



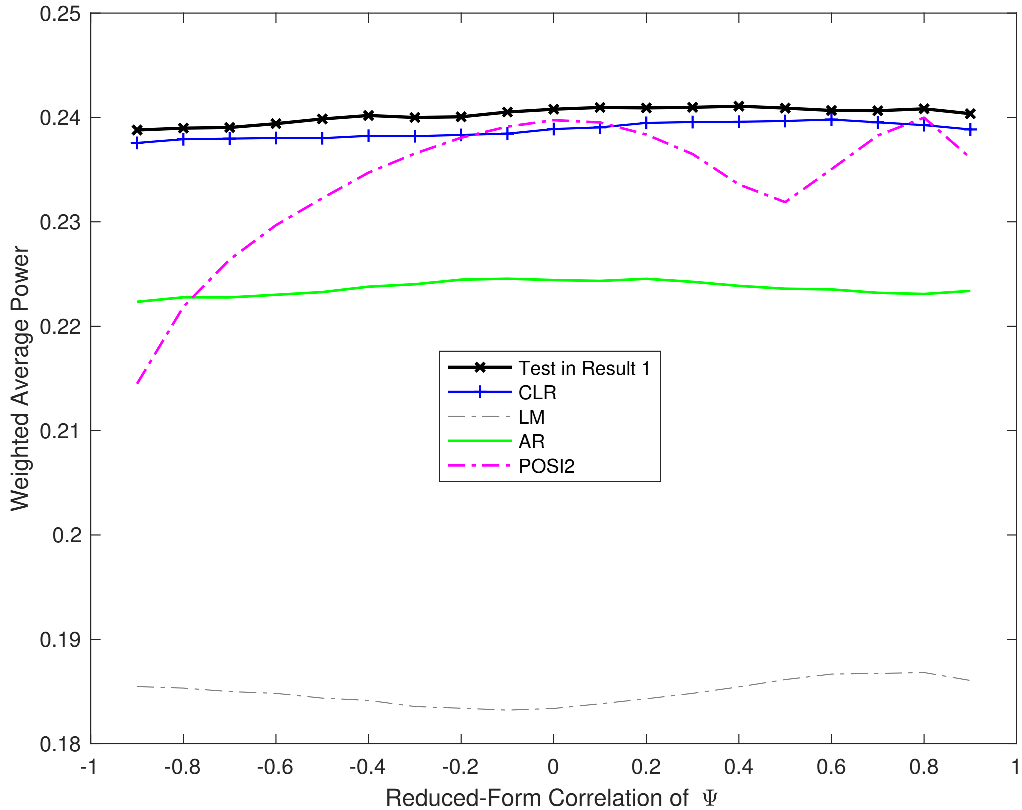
(c) $\lambda = 10$



(d) $\lambda = 20$

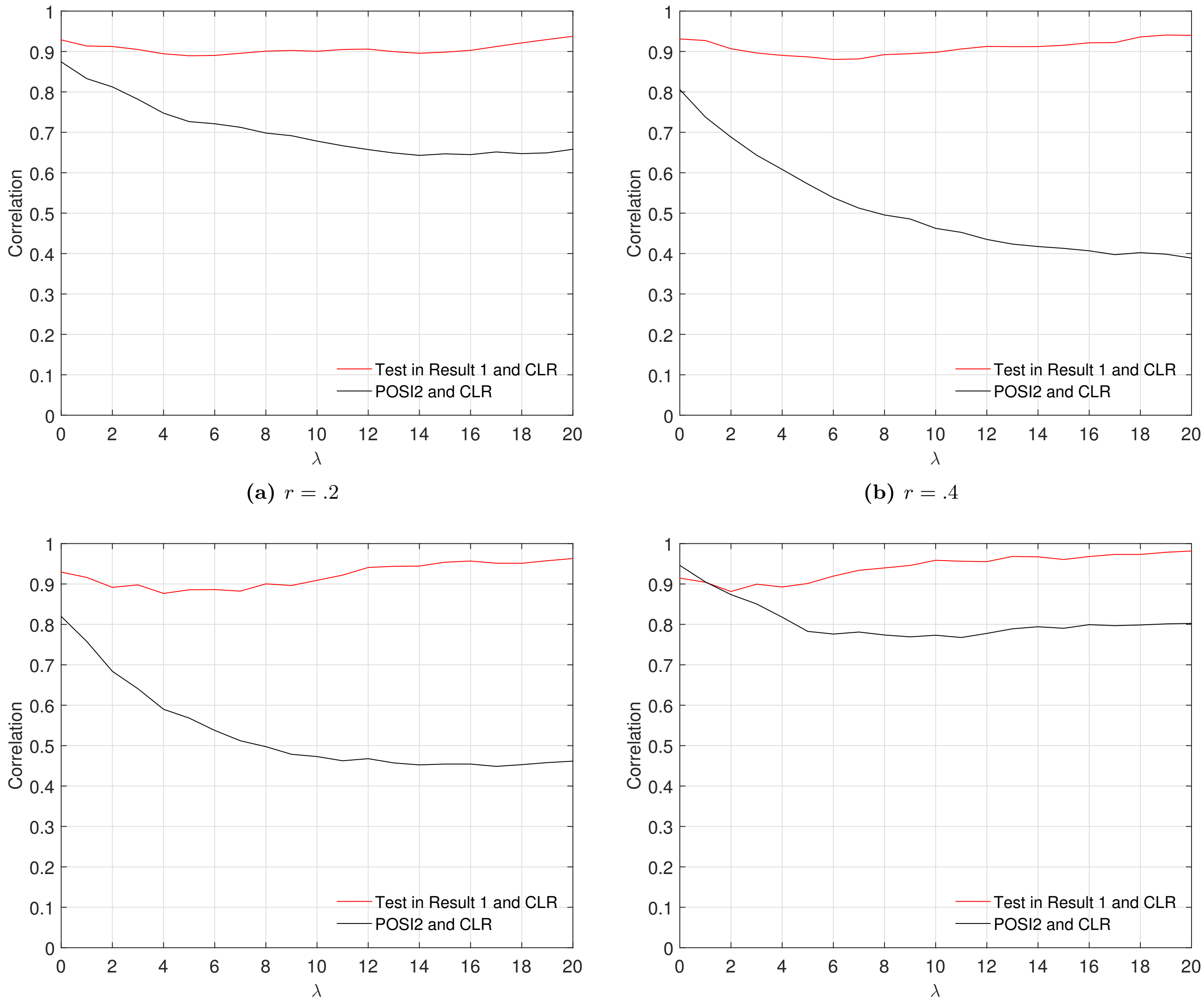
Description: Figure 2 reports power curves for WAP-similar test in Result 1 (denoted R1), the CLR, LM, AR, the POSI2 in Andrews et al. (2006), and an infeasible test that rejects whenever $|\Pi'\Phi^{-1/2}S|/(\Pi'\Phi^{-1}\Pi)^{1/2} > 1.96$. The POSI2 test is evaluated at $\beta = 2.1$ and $\lambda = 1$. The figure suggests that the power curves of the test R1 and that of the CLR are almost indistinguishable. The numbers in the box are weighted average power for each fixed λ and a uniform grid of 121 points for $\lambda^{1/2}(\beta - \beta_0) \in [-6, 6]$. The script used to generate this figure is `PowerPlots.m`.

Figure 3: WAP-similar test (R1) vs. CLR, LM, AR (WAP)
($k = 4,\ \Phi = \mathbb{I}_4,\ \Psi(1,1) = \Psi(2,2) = 1$)



Description: Figure 3 presents a comparison of weighted average power between the test in Result 1 (denoted R1), the CLR, LM, AR, and the POSI2 in Andrews et al. (2006). The weights in (3.2) for the parameters (β, Π) are evaluated using $\Phi = \mathbb{I}_4$ and $\Psi = [1, r; r, 1]$, with $r \in [-.9 : .1 : .9]$. The null hypothesis is $\beta_0 = 0$. The sample size is $n = 100$. The figures use 500 draws from (β, Π) and 1,000 Monte-Carlo draws to compute the power at each point. The POSI2 test is evaluated at $\beta = 2.1$ and $\lambda = 1$. The figure suggests that the WAP of R1 and that of the CLR are almost the same. The script used to generate this figure is `WAPComparison_KronIV.m`.

Figure 4: Correlation between the WAP-similar test and the CLR
($k = 4$)



(a) $r = .2$

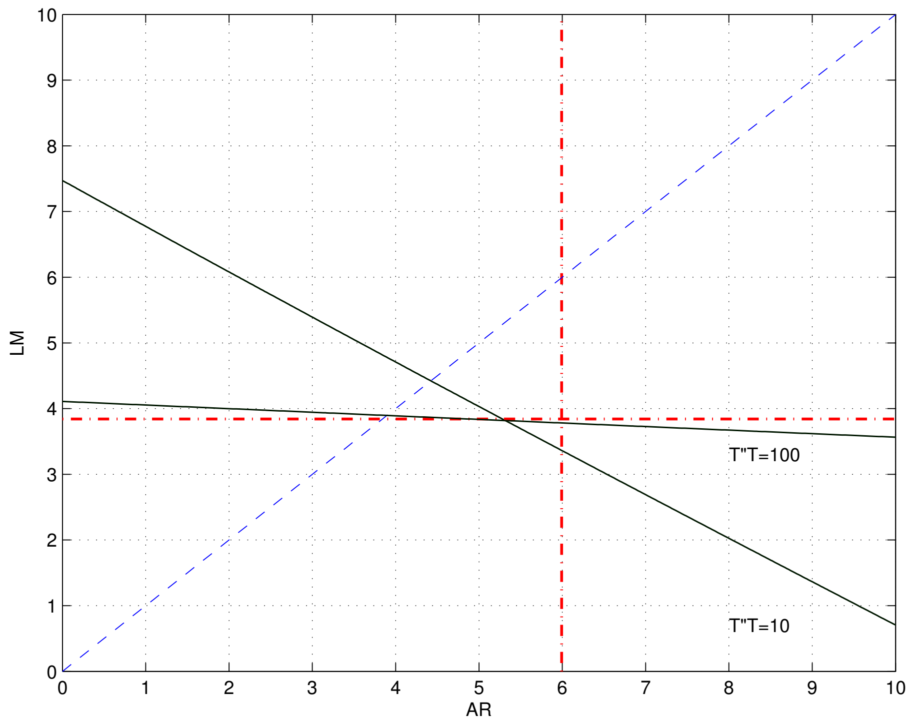
(b) $r = .4$

(c) $r = .6$

(d) $r = .8$

Description: Figure 4 reports the correlation (under the null hypothesis) between the WAP-similar test in Result 1 and both the CLR and the POSI2 in Andrews et al. (2006). The design under consideration is almost the same as in Figure 1. The null hypothesis is $\beta_0 = 0$. The matrix Ψ is assumed to have unit diagonal elements and correlation parameter $r \in \{.2, .4, .6, .8\}$. The matrix $\Phi = \mathbb{I}_k$. The number of instruments is $k = 4$. It has been argued that the power of the CLR is close to that of POSI2 and we include it for comparison. The POSI2 test requires to be evaluated a pair (β, λ). We consider the alternative $\beta = 2.1$ and $\lambda = 1$. The alternative for β is chosen to be close to $1/r = 2$. Even though the rate of Type I error of these tests does not depend on the concentration parameter, their correlation changes with λ. We use a uniform grid for $\lambda \in [0, 20]$ of 20 points. The script used to generate this figure is `CorrelationWAPvsCLR.m`.

Figure 5: 5% Conditional Critical Region
(AR,LM), $k = 2$



(Blue, dashed) Boundary of the sample space: AR $\geq$ LM, where AR$\equiv$ $S'S$ and LM$\equiv$ $(S'T)^2/(T'T)$. (Red, dot-dashed) 5% critical values for the AR and the LM statistics obtained as the upper 5% quantiles of the distributions χ^2_2 and χ^2_1, respectively. The conditional critical region is the collection of (AR, LM) points at the right of the black (solid) lines (large AR and large LM). Each solid line traces the boundary of the rejection region of the WAP-similar test for a given value of $T'T \in \{10, 100\}$. The command `ezplot` in Matlab is used to graph the solution to the equation z(AR, LM, $T'T$)-$c(T'T;\alpha) = 0$.

REFERENCES

Abramowitz, M. and I. Stegun (1964): *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, vol. 55, Dover publications, New York.

Aliprantis, C. and K. Border (2006): *Infinite Dimensional Analysis: a Hitchhiker's Guide*, Springer Verlag, 3rd ed.

Anderson, T. and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 20, 46–63.

Andrews, D., M. Moreira, and J. Stock (2004): "Optimal Invariant Similar Tests for Instrumental Variables Regression," *Working paper, Harvard University.*

——— (2006): "Optimal Two-sided Invariant Similar Tests for Instrumental Variables Regression," *Econometrica*, 74, 715–52.

Andrews, I. (2016): "Conditional Linear Combination Tests for Weakly Identified Models," *Econometrica*, 2155–2182.

Billingsley, P. (1995): *Probability and Measure*, John Wiley & Sons, New York, 3rd ed.

Brown, L. D. (1986): "Fundamentals of statistical exponential families with applications in statistical decision theory," *Lecture Notes-monograph series*, 9, i–279.

Chamberlain, G. (2007): "Decision Theory Applied to an Instrumental Variables Model," *Econometrica*, 75, 609–652.

Chernozhukov, V., C. Hansen, and M. Jansson (2009): "Admissible Invariant Similar Tests for Instrumental Variables Regression," *Econometric Theory*, 25, 806–818.

Dudley, R. (2002): *Real Analysis and Probability*, vol. 74, Cambridge University Press.

Ferguson, T. (1967): *Mathematical Statistics: A Decision Theoretic Approach*, vol. 7, Academic Press New York.

Le Cam, L. (1986): *Asymptotic Methods in Statistical Decision Theory*, Springer Verlag.

Lehmann, E. and J. Romano (2005): *Testing Statistical Hypotheses*, Springer Texts in Statistics, Springer Verlag.

Lehmann, E. L. and G. Casella (1998): *Theory of point estimation*, vol. 31 of *Springer Texts in Statistics*, Springer, New York, 2nd ed.

Linnik, J. (1968): *Statistical Problems with Nuisance Parameters*, vol. 20, American Mathematical Society.

Mardia, K. and P. Jupp (2000): *Directional Statistics*, Wiley Series in Probability and Statistics, Wiley.

Moreira, M. (2003): "A conditional likelihood ratio test for structural models," *Econometrica*, 71, 1027–1048.

Moreira, M. and H. Moreira (2015): "Optimal Two-Sided Tests for Instrumental Variables Regression with Heteroskedastic and Autocorrelated Errors," Working Paper, EPFGV.

Müller, U. K. (2011): "Efficient Tests under a Weak Convergence Assumption," *Econometrica*, 79, 395–435.

Neyman, J. (1935): "Sur la Vérification des Hypothèses Statistiques Composées," *Bulletin de la Société Mathématique de France*, 346–366.

Ok, E. (2019): *Probability theory with economic applications*, forthcoming.

Olver, F. W. (1997): *Asymptotics and special functions*, AKP Classics, Academic press, 2nd ed.

Phillips, P. C. (1989): "Partially identified econometric models," *Econometric Theory*, 5, 181–240.

Rudin, W. (2005): *Functional Analysis.*, International Series in Pure and Applied Mathematics, McGraw-Hill, New York.

——— (2006): *Real and Complex Analysis*, Tata McGraw-Hill Education, 3rd ed.

Staiger, D. and J. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stein, E. (2011): *Functional Analysis: Introduction to Further Topics in Analysis*, Princeton Univ Press.

Stroock, D. (1999): *A Concise Introduction to the Theory of Integration*, Birkhauser, 2nd ed.

Van der Vaart, A. (2000): *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Van der Vaart, A. and J. Wellner (1996): *Weak Convergence and Empirical Processes.*, Springer, New York.

Wald, A. (1950): *Statistical Decision Functions*, Oxford, England: Wiley.

Yogo, M. (2004): "Estimating the Elasticity of Intertemporal Substitution when Instruments are Weak," *Review of Economics and Statistics*, 86, 797–810.

APPENDIX A: MAIN TEXT APPENDIX.

A.1. ***Proof of Theorem 1***

Let $\mathbf{X}$ denote the sample space, $\boldsymbol{\Theta}$ the parameter space, f the statistical model, and Θ_0 the null hypothesis. Assume Θ_0 is nonempty, closed, and such that its topological boundary, $\mathrm{Bd}(\Theta_0)$, equals Θ_0.

Let $\mathcal{C}$ denote the class of all tests; that is, all measurable functions ϕ such that $\phi(x) \in [0,1]$ for almost every $x \in \mathbf{X}$. Topologize this space using the weak* topology on L^∞ (see Section B.1.1 of the Online Supplementary Material for details). Denote the resulting topological space as $(\mathcal{C}, \mathcal{T}^*_{\mathcal{C}})$.

The first lemma of this appendix shows that the set of α-similar tests is compact.

Lemma 1: *The set*

$$\mathcal{C}(\alpha\text{-}s) \equiv \left\{ \phi \in \mathcal{C} \;\middle|\; \mathbb{E}_\theta[\phi(X)] - \alpha] \equiv \int_{\mathbf{X}} (\phi(x) - \alpha) f(x;\theta) = 0 \quad \forall \theta \in Bd(\Theta_0) \right\}$$

is compact relative to $(\mathcal{C}, \mathcal{T}^*_{\mathcal{C}})$.

Proof: See Section B.1.2 of the Online Supplementary Material. *Q.E.D.*

Lemma 1 is used to prove part ii) of Theorem 1. The weak* compactness of $\mathcal{C}(\alpha\text{-s})$ will allow the application of an essentially complete class Theorem [See Theorem 3, p. 87, Chapter 2 in Ferguson (1967)].

The second lemma in this appendix shows that the test that minimizes weighted average risk subject to an α-similarity constraint is well defined and, more importantly, admissible relative to all tests.

Lemma 2: *Let w denote a full-support probability measure over Θ_1 with p.d.f. p. Define the* minimum average risk *over the set of α-similar procedures as*

$$M(w) \equiv \underset{\phi \in \mathcal{C}(\alpha\text{-}s)}{\arg\min} \int_{\Theta_1} R(\phi, \theta) dw(\theta) \ ,$$

and suppose that Assumption F0 holds. Then, $M(w)$ is a nonempty set and each $\phi^* \in M(w)$ is admissible in $\mathcal{C}(\alpha\text{-}s)$ and $\mathcal{C}$.

Proof: See Section B.1.3 of the Online Supplementary Material. *Q.E.D.*

Lemma 2 is used to prove part i) of Theorem 1.

Proof of Theorem 1: We use the previous lemmas to prove the theorem.

Part i: Let $\boldsymbol{\phi}^{w,\alpha}_{\text{WAP}}$ be a ω-WAP-similar test of level α and let $\mathcal{C}(\alpha\text{-}s)$ be the class of α-similar tests as defined in Lemma 1. By definition, $\boldsymbol{\phi}^{w,\alpha}_{\text{WAP}} \in \mathcal{C}(\alpha\text{-}s)$ and

$$\begin{aligned}
\text{WAP}(\boldsymbol{\phi}^{w,\alpha}_{\text{WAP}}, w) \quad &\equiv \quad \int_{\Theta_1} \Big(\int_{\mathbf{X}} \boldsymbol{\phi}^{w,\alpha}_{\text{WAP}}(x) f(x;\theta) dx \Big) dw(\theta), \\
&= \quad \int_{\Theta_1} \mathbb{E}_\theta \left[\boldsymbol{\phi}^{w,\alpha}_{\text{WAP}}(x) \right] dw(\theta), \\
&= \quad 1 - \int_{\Theta_1} R(\boldsymbol{\phi}^{w,\alpha}_{\text{WAP}}, \theta) dw(\theta), \\
&\qquad \text{(by the definition of risk function)} \\
&\geq \quad \text{WAP}(\boldsymbol{\phi}, w), \quad \forall \boldsymbol{\phi} \in \mathcal{C}(\alpha\text{-}s).
\end{aligned}$$

This implies

$$\int_{\Theta_1} R(\boldsymbol{\phi}^{w,\alpha}_{\text{WAP}}, \theta) dw(\theta) \leq \int_{\Theta_1} R(\boldsymbol{\phi}, \theta) dw(\theta), \quad \forall \boldsymbol{\phi} \in \mathcal{C}(\alpha\text{-}s).$$

Implying WAP-similar tests of level α are average risk minimizers subject to an α-similarity constraint. Lemma 2 thus implies that $\boldsymbol{\phi}^{w,\alpha}_{\text{WAP}}$ is admissible in the class of all tests, $\mathcal{C}$, provided Assumption F0 holds.

Part ii: The proof is based on the essentially complete class theorem; see Theorem 2.9.2 and 2.10.3 in Ferguson (1967) and also Le Cam (1986), Chapter 2, Theorem 1.

Note first that the class $\mathcal{C}(\alpha - s)$ is essentially complete relative to itself (as it contains all the α-similar tests). Note that the set $\mathcal{C}(\alpha - \text{s})$ is weak* compact by Lemma 1. In addition, the risk function of the testing problem $R(\phi, \theta)$ is—by definition of weak* topology—continuous (in ϕ) for all $\theta \in \Theta$. This verifies the assumptions of Theorem 2.10.3, p. 87, in Ferguson (1967).

Following Definition 3 Ferguson (1967) p. 50, $\phi^* \in \mathcal{C}(\alpha - s)$ is said to be an extended Bayes test if for every $\epsilon > 0$ there is a prior distribution $w_\epsilon(\theta)$ such that:

$$\int_{\Theta_1} R(\phi^*, \theta) dw_\epsilon(\theta) \leq \int_{\Theta_1} R(\boldsymbol{\phi}^{w_\epsilon,\alpha}_{\text{WAP}}, \theta) dw_\epsilon(\theta) + \epsilon.$$

Theorem 2.10.3 in Ferguson (1967) implies that the set of *extended Bayes tests* in $\mathcal{C}(\alpha - \text{s})$ is *essentially complete*. This essential completeness means that for any other test $\phi \in \mathcal{C}$ there is a test ϕ^* extended Bayes in $\mathcal{C}(\alpha\text{-s})$ such that:

$$R(\phi^*, \theta) \leq R(\phi, \theta)$$

for all θ. Since ϕ is admissible and α-similar $R(\phi^*, \theta) \leq R(\phi, \theta)$ for all θ implies that $R(\phi^*, \theta) = R(\phi, \theta)$. Therefore, any admissible, α-similar test is risk equivalent to an extended Bayes test. This implies that for any $\epsilon > 0$ there is a probability measure w_ϵ such that

$$\text{WAP}(\phi, w_\epsilon) = \text{WAP}(\phi^*, w_\epsilon) \geq \text{WAP}(\boldsymbol{\phi}^{w_\epsilon,\alpha}_{\text{WAP}}, w_\epsilon) - \epsilon$$

Consequently, any admissible, α-similar test is an extended WAP-similar test of level α.

Q.E.D.

A.2. ***Proof of Result 1***

The statistical model under consideration is

$$\text{(A.1)} \qquad \widehat{\gamma}_n \sim \mathcal{N}_{2k}\left(\begin{pmatrix} \beta\Pi \\ \Pi \end{pmatrix}, \frac{\Sigma}{n} \right), \quad \text{where } \widehat{\gamma}_n \equiv \begin{pmatrix} (Z'Z)^{-1}Z'y \\ (Z'Z)^{-1}Z'x \end{pmatrix}.$$

As explained in footnote 12 in the main body of the paper, if we assume $\Sigma = \Psi \otimes \Phi$ the model can be reparameterized as

$$\text{(A.2)} \qquad \begin{pmatrix} S_n \\ T_n \end{pmatrix} \sim \mathcal{N}_{2k}\left(\rho\left(\phi \otimes \omega\right) \,,\, \mathbb{I}_{2k}\right), \quad \rho \in \mathbb{R}_+, \quad \phi \in \mathcal{S}^1, \quad \omega \in \mathcal{S}^{k-1}.$$

The test in Result 1 uses the likelihood of (A.2), denoted $f(S, T; \rho, \phi, \omega)$. The following lemmas provide expressions for the corresponding integrated likelihood using the independent distributions:

$$\phi \sim \mathcal{U}(\mathcal{S}^1), \; \omega \sim \mathcal{U}(\mathcal{S}^{k-1}), \; \rho \sim \sqrt{\chi^2_k}.$$

The uniform measure over the $r-1$ dimensional sphere $\mathcal{S}^{r-1}$ is defined in Stroock (1999) and denoted $\lambda_{\mathcal{S}^{r-1}}(\cdot)$.

The following lemmas derive an expression for the integral of $f(S,T;\rho,\phi,\omega)$ with respect to, first, the distributions of ρ and ω and, second, the distribution of ϕ.

Lemma 3: *Let m denote the density of ρ. Then*

$$\int_{\mathbb{R}^+}\left(\int_{\mathcal{S}^{k-1}} f(S,T;\rho,\phi,\omega)d\lambda_{\mathcal{S}^{k-1}}(\omega)\right)m(\rho)d\rho \propto \exp\Big(-\frac{1}{2}[S'S+T'T]\Big)\exp\Big(\frac{1}{4}\phi'Q\phi\Big),$$

where $Q\equiv[S,T]'[S,T]$.

Proof: See Section B.2.1 of the Online Supplementary Material. *Q.E.D.*

Lemma 4:

$$\int_{\mathcal{S}^1}\exp\Big(\frac{1}{4}\phi'Q\phi\Big)d\lambda_{\mathcal{S}^1}(\phi)\propto\exp\Big(\frac{1}{8}\Big(\zeta_{\max}+\zeta_{\min}\Big)\Big)I_0\left(\frac{1}{8}(\zeta_{max}-\zeta_{min})\right),$$

where

$$\zeta_{max}=\frac{1}{2}\left[(S'S+T'T)+\sqrt{(S'S-T'T)^2+4(S'T)^2}\right],$$

$$\zeta_{min}=\frac{1}{2}\left[(S'S+T'T)-\sqrt{(S'S-T'T)^2+4(S'T)^2}\right],$$

are the largest and smallest eigenvalues of the matrix Q and $I_0(\cdot)$ is the modified Bessel function of the first kind, defined in Abramowitz and Stegun (1964), Section 9.6, p. 375.

Proof: See Section B.2.2 of the Online Supplementary Material. *Q.E.D.*

Proof of Result 1: We now derive the WAP-similar test. From the definitions of ζ_{max} and ζ_{min}:

$$\frac{1}{8}\Big(\zeta_{max}+\zeta_{min}\Big)=\frac{1}{8}\Big(S'S+T'T\Big),\quad \frac{1}{8}\Big(\zeta_{max}-\zeta_{min}\Big)=\frac{1}{8}\Big((S'S-T'T)^2+4(S'T)^2\Big)^{1/2}.$$

From Lemma 3 and Lemma 4 it follows that the integrated likelihood, denoted $f^*(S,T)$ with respect to the independent weights

$$\phi\sim\mathcal{U}(\mathcal{S}^1)\quad \omega\sim\mathcal{U}(\mathcal{S}^k)\quad \rho\sim\sqrt{\chi^2_k}$$

is:

$$f^*(S,T)\propto\exp(-\frac{1}{2}[S'S+T'T])\exp\left(\frac{1}{8}\Big(S'S+T'T\Big)\right)I_0\left(\frac{1}{8}\Big((S'S-T'T)^2+4(S'T)^2\Big)^{1/2}\right).$$

The denominator in the expression of the WAP-similar test is

$$f(S|T;\beta_0)\propto\exp\Big(-\frac{1}{2}S'S\Big).$$

Consequently:

$$z_{\text{WAP}}(S,T)\propto\exp\Big(-\frac{1}{2}T'T\Big)\exp\Big(\frac{1}{8}[S'S+T'T]\Big)I_0\left(\frac{1}{8}\Big((S'S-T'T)^2+4(S'T)^2\Big)^{1/2}\right).$$

The quantile function $c(T,\alpha)$ is continuous in T and, therefore, measurable. So that the WAP-similar test rejects if and only if the test statistic

$$S'S - T'T + 8\ln\left[I_0\left(\frac{1}{8}\left((S'S - T'T)^2 + 4(S'T)^2\right)^{1/2}\right)\right]$$

is larger than the critical value function $c^*(T,\alpha)$, defined as the $1-\alpha$ quantile (conditional on T) of the expression above under the distribution $S \sim \mathcal{N}_k(\mathbf{0}, \mathbb{I}_k)$. This test is equivalent to the one presented in Result 1.