Optimal Conditional Inference for Instrumental Variables Regression*

Laura Chioda and Michael Jansson

May 11th, 2005

Abstract

In the context of the instrumental variables model we focus on tests that are similar conditional on the first stage F statistic. We argue that in some economic applications, it is desirable to conduct inference conditionally on the first stage F statistic. Assuming homoskedastic Gaussian errors and known covariance matrix, we derive the power envelopes for conditionally similar and conditionally unbiased tests for the coefficient on the endogenous regressor. Making use of Staiger and Stock (1997) asymptotics, we show that the efficiency bounds derived under the assumptions of Gaussian errors and known covariance matrix can also be attained in large samples when the reduced form covariance matrix has to be estimated and the errors are nonnormal. Relying on the theory of limits of experiments, we obtain Gaussian asymptotic efficiency bounds for conditionally similar and conditionally unbiased tests. A Monte Carlo study is conducted to assess the performance of the conditional testing procedures.

*Corresponding author Laura Chioda, Department of Economics, 549 Evans Hall, #3880, Berkeley CA 94720-3880, email: lchioda@econ.berkeley.edu. We would like to thank Jim Powell, Tom Rothenberg and Paul Ruud for extensive conversations and valuable suggestions. We are grateful to Don Andrews, Peter Bickel, Ken Chay, Victor Chernozhukov, Jerry Hausman, Bo Honore, Guido Imbens, Paco Martorell, Marcelo Moreira, Ulrich Müller, Whitney Newey, Jack Porter, Jim Stock, Tom Stoker, Mark Watson and seminar participants at Berkeley, Caltech, Boston College, Harvard, MIT, Princeton, University of Michigan, University of Pittsburgh, UC Santa Barbara, University of Wisconsin-Madison for helpful comments.

1 Introduction

This paper focuses on optimal inference in an instrumental variables framework when instruments can be arbitrarily weak. Much of the recent theoretical econometric work on this topic has found helpful motivation in Angrist and Krueger's (1991) study on the return to schooling. More generally, weak instruments are quite common in economics (e.g. see references in Zivot, Startz, and Nelson (1998)) and represent a potential problem for any application that relies on instrumental variables methods.[1]

Identification in the IV model is achieved by assuming the existence of instrumental variables uncorrelated with the errors. Even when such instruments exist, inference on β – the coefficient on the endogenous regressor – can be problematic when there is low correlation between excluded instruments and endogenous variables.[2] In the presence of weak instruments, conventional asymptotics fail to provide good approximations regardless of the sample size (Bound, Jaeger, and Baker (1995), Nelson and Startz (1990) and Rothenberg (1984)). As a consequence, the standard approach of building Wald type test statistics around estimators, which is founded on conventional asymptotics, gives rise to procedures which exhibit significant size distortions.[3]

Most contributions have focused on developing asymptotically valid methods for testing hypotheses about the coefficient on the endogenous regressor in the presence of weak identification.[4] With the exception of Moreira (2005), who establishes optimality of the Anderson-Rubin (1949) test in the just identified case, and Andrews, Moreira, and Stock (2004), hereafter AMS, who derive point optimal invariant similar tests, none of the existing testing procedures is known to enjoy optimality properties when the instruments might be weak. In this paper, we derive optimal (conditionally similar) tests for the coefficient of the endogenous regressor that are robust to weak instruments.

We consider two-sided testing problems concerning a simple null hypothesis on β when the number of instruments is greater than or equal to two. Our analysis proceeds under a rotation invariance restriction previously considered by AMS, Chamberlain (2001), Chamberlain and Imbens (2004), and Hillier (2004). The invariance restriction guarantees that inference concerning β is unaffected by the order in which the instrumental variables appear.[5] In addition to imposing invariance, we focus our attention on tests that are conditionally similar in the sense that they have correct size conditional on the first stage F statistic.

[1] For a review of the pervasive role of IV techniques in the applied economics, see Angrist and Krueger (2001).

[2] Inference in the IV model may also be problematic if there are a large number of overidentifying restrictions (Bekker (1994), Chao and Swanson (2003, 2005), Donald and Newey (2001), Hahn (2002), Hansen, Hausman, and Newey (2004), and Stock and Yogo (2003)). In this paper we consider only Staiger and Stock (1997) asymptotics, postponing to further research the analysis of inference under many instruments asymptotics.

[3] Building on the work of Gleser and Hwang (1987), Dufour (1997) shows that confidence regions based on Wald type statistics do not have correct coverage probabilities. Wang and Zivot (1998) show that the standard likelihood ratio test does not have correct size.

[4] See for instance, Staiger and Stock (1997), Zivot, Startz, and Nelson (1998), Wang and Zivot (1998), Dufour and Jasiak (2001), Kleibergen (2002), Dufour and Taamouti (2005), Moreira (2003), Startz, Zivot, and Nelson (2003). For reviews, see Dufour (2003) and Stock, Yogo, and Wright (2002).

[5] As noted by AMS, the invariance restriction does not rule out those tests commonly used in practice, such as the Anderson-Rubin test, the Conditional Likelihood Ratio (CLR) and Wald tests (Moreira (2003)), and the Lagrangian Multiplier (LM) test (Kleibergen (2002) and Moreira (2001)).

Whereas the restriction to invariant testing procedures is conventional, the decision to condition on the first stage F statistic requires motivation/justification.[6] In the IV model, the first stage F statistic is a specific ancillary for the parameter of interest; that is, its distribution does not depend on β. It is therefore tempting to condition on it. As in Jansson and Moreira (2004), in the context of nearly integrated regressors, this conditional approach is motivated by a generalization of the notion of ancillarity and by an extension of the conditionality principle (Cox and Hinkley (1974)). Unlike Jansson and Moreira (2004), we provide an detailed discussion of the merits of the conditional approach and our argument for conditioning on the first stage F statistic goes beyond the recognition that it is a specific ancillary. We argue that there is a precise sense in which conditioning on the first stage F statistic is analogous to the standard approach of treating the observed regressors as fixed in an ordinary regression problem. In the case of the linear regression with random regressors, carrying out inference conditionally or unconditionally delivers the same conclusions (asymptotically). As we show below, the same is not true in the IV model when the identification is weak. Therefore, it is important to understand when the conditional point of view should be adopted. We argue that the desirability of the conditional approach depends on the source of endogeneity underlying the IV model and discuss specific examples to illustrate this point.

As in most other applications of conditional inference, our main motivation for conditioning on the first stage F statistic is the desire to condition on the experiment actually conducted, thereby making the probability calculations more relevant to the data under study (Cox (1988)). An alternative motivation for employing one of the testing procedures developed herein is the desire to avoid pretest bias. Several authors (e.g., Hansen, Hausman, and Newey (2004), Staiger and Stock (1997), Stock and Yogo (2002), Wang and Zivot (1998), and Zivot, Startz, and Nelson (1998)) have highlighted the importance of the first stage F statistic in detecting weak instruments problems and have argued in favor of pretesting procedures based on this statistic. By construction, the conditionally similar tests studied here take into consideration the realized value of the first stage F statistic and eliminate potential biases induced by "taking a peak" at the first stage F statistic before doing inference on β.

Following Anderson, Kunitomo, and Sawa (1982) and Rothenberg (1984), we study a model with exogenous instruments, homoskedastic Gaussian errors, and known reduced form covariance matrix.[7] Under these assumptions, we obtain the power envelope for (invariant) conditionally similar tests, and show that no uniformly most powerful test exists within the class of (invariant) conditionally similar testing procedures. Moreover, exploiting the analytic description of the efficiency bound, we propose a feasible test (labelled CAR) that matches the power envelope over a large portion of the parameter space. In addition, by strengthening conditional similarity with conditional unbiasedness, and therefore focusing only on conditionally similar tests whose power under the alternative is no lower than under

[6] In an estimation context, Forchini and Hillier (2003) emphasize the importance of adjusting standard errors, and other measures of the precision of estimates, to account for the degree of identification in the sample. Because the first stage F statistic represents a natural measure of the degree of identification, they conclude that conditioning on it is desirable. As will become clear in the sequel, our argument for conditioning on the first stage F statistic is somewhat different from that of Forchini and Hillier (2003).

[7] The same framework has also been adopted by Moreira (2001, 2003a), and AMS.

the null, we obtain a uniformly most powerful conditionally unbiased (UMPCU) test.

Our conditional optimality results have a twofold interpretation. In those instances in which conditioning on the specific ancillary provides a more accurate description of the data underlying the experiment, our findings contribute to a theory of optimal hypothesis testing for conditionally (similar) tests. Alternatively, if the unconditional point of view is more appropiate then imposing the conditional similarity restriction can be viewed as an alternative way of obtaining similar tests, and the optimality results inform us how to maximize the unconditional power subject to the conditional similarity restriction. (As discussed previously, the conditional similarity restriction is natural from the unconditional point of view whenever the analyst wants to avoid pretest bias caused by "taking a peak" at the first stage F statistic before doing inference on β.)

The assumptions on the disturbances and on the covariance matrix may appear restrictive at first. However, they simply serve the purpose of specifying the finite sample model in such a way that it enjoys the same statistical properties as the limiting experiment associated with a model with unknown covariance matrix and possibly nonnormal errors under weak instruments asymptotics. Indeed, relying on Staiger and Stock (1997) asymptotics and making use of Le Cam's limit of experiments theory (Le Cam and Yang (2000) and van der Vaart (2002)), we show that the efficiency bounds derived under the assumptions of Gaussian errors and known covariance matrix can be attained in large samples when the reduced form covariance matrix has to be estimated and the errors are nonnormal. The derivation of this result exploits the fact that the Gaussian model with known covariance matrix and fixed instruments captures those statistical features of the limiting experiment that are essential to the derivation of our results.

Our analytical (conditional) optimality results are supplemented with a numerical study. First, adopting the unconditional point of view, we quantify the efficiency loss corresponding to the conditional similarity restriction. Conditioning on the first stage F statistic is shown to result in a loss of unconditional power, especially for small numbers of instruments and low degree of correlation between the error terms of the two reduced form equations. For very weak instruments, on the other hand, the discrepancy between unconditional power curves and the power envelope for conditionally similar tests shrinks and the latter is not uniformly dominated. Adopting the conditional point of view, we then show that the loss of unconditional power can be explained in part by the fact that conventional testing procedures (e.g., Anderson-Rubin (1949), Kleibergen's (2002) LM test, and Moreira's (2003) CLR test) trade off conditional size distortions for unconditional power by (implicitly) choosing the critical values as a function of the first stage F statistic. Finally, we assess the performances of the CAR and the UMPCU tests relative to the efficiency bound for conditionally similar tests. The shortcoming of the CAR test becomes visible only for large values of the correlation coefficient between the error terms of the two reduced form equations. For those same values of the reduced form correlation coefficient, the power of the CAR test tends to dip below the significance level over a small range of values of β. By definition, the UMPCU test does not suffer from this problem, but our results show that imposing unbiasedness introduces a trade-off between the loss of power induced by the additional restriction and the bias of the CAR test.

The remainder of the paper is organized as follows. Section 2 introduces the finite sample Gaussian model and the invariance restriction, providing details on the corresponding maximal invariant. Sections 3 and 4 present the optimality results for conditionally similar and for conditionally unbiased tests, respectively. In Section 5, we motivate the conditional approach. (The discussion is supplemented by Appendix C.) Section 6 extends the results by augmenting the model with exogenous regressors, relaxing the assumptions of normal errors and known covariance matrix, and deriving asymptotic Gaussian power envelopes under weak instruments asymptotics. Finally, Section 7 presents numerical results, while Section 8 concludes and outlines directions for further research.

2 Preliminaries

2.1 The model

We consider an overidentified Gaussian model with one endogenous variable, no exogenous variables, and k nonrandom instruments.[8] The model consists of a structural equation and a reduced form equation given by

$$\begin{aligned} y_1 &= y_2\beta + u, \quad (1)\\ y_2 &= Z\Pi + v_2, \quad (2) \end{aligned}$$

where y_1 and y_2 are n-dimensional vectors of endogenous variables, Z is an $n \times k$ full column rank matrix of nonrandom instruments, and each row of the $n \times 2$ matrix of disturbances $[u : v_2]$ is assumed to be iid (bivariate) normal with mean zero and a nonsingular covariance matrix

$$\Sigma = E\left[(u_{i1}, v_{i2})'(u_{i1}, v_{i2})\right] = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}. \quad (3)$$

The scalar β is the parameter of interest and Π is a k-dimensional vector of nuisance parameters. The reduced form equations, relating the endogenous variables to the instruments, can be written as

$$\begin{aligned} y_1 &= Z\Pi\beta + v_1, \\ y_2 &= Z\Pi + v_2, \end{aligned}$$

where $v_1 = u + \beta v_2$. As in Moreira (2005), the covariance matrix for the i-th row of $[v_1 : v_2]$,

$$\Omega = E\left[(v_{i1}, v_{i2})'(v_{i1}, v_{i2})\right] = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix}, \quad (4)$$

[8] In Section 6.1, we extend the model by introducing exogenous regressors, while in Section 6.2 we relax the assumptions of Gaussian errors (and known covariance matrix).

is assumed to be known and nondiagonal.[9]

We study testing problems characterized by a null hypothesis of the form $\beta = \beta_0$. Without loss of generality, the value hypothesized under the null is normalized to be equal to zero.[10] We are interested in conducting (optimal) inference on β in the presence of the unknown k-dimensional vector of nuisance parameters Π.

We focus on conditionally similar tests, i.e. similar tests conditional on the first stage F statistic, when k is greater than or equal to two. In fact, in the just identified case, conditioning on the first stage F statistic is equivalent to conditioning on y_2, in which case β is no longer identified. More importantly, the null hypothesis $\beta = 0$ cannot be expressed as a restriction on the moments of the conditional distribution of $Z'y_1$ given $Z'y_2$.

In the overidentified case, inference concerning β is nontrivial. The probability model for $Y = [y_1 : y_2]$ constitutes a curved exponential family (a minimal sufficient statistic for this model is the $k \times 2$ matrix $Z'Y$, while the parameter $(\beta, \Pi')'$ is of dimension $k + 1$), and the curvature of the model rules out the possibility of applying conventional results for exponential families, such as those in Lehmann (1997). In addition, Theorem 2(c) in Moreira (2005) implies that no Uniformly Most Powerful (UMP) unbiased test exists in the overidentified case.[11]

2.2 Reduction by Invariance and Conditioning on Specific Ancillaries

The analysis builds on the invariance result considered by AMS, Chamberlain (2001), Chamberlain and Imbens (2004), and Hillier (2004). The problem of testing $H_0 : \beta = 0$ against the two sided (or one-sided) alternative is invariant under full-rank orthogonal transformations of the matrix of instruments Z.[12] Sufficiency and invariance considerations enable AMS to reduce the data to a maximal invariant of dimension three and to shrink the dimension of the parameter space to two. In particular, the k-dimensional nuisance parameter Π is reduced to a scalar nuisance parameter given by

$$\lambda_F = \omega_{22}^{-1} \Pi' Z' Z \Pi,$$

also known as the concentration coefficient, a unitless measure of the strength of the instruments (Rothenberg (1984) and Stock, Yogo, and Wright (2002)).

The discrepancy between the dimension of the parameter space and that of the data implies that

[9] Previous authors (e.g. Anderson, Kunitomo, and Sawa (1982) and Rothenberg (1984)) have emphasized that the simultaneous equations model with known reduced form covariance matrix has a simpler mathematical structure than the model with unknown covariance matrix, but test procedures for the two models behave similarly in moderately sized samples.

[10] If the null hypothesis is of the form $\beta = \beta_0$, with $\beta_0 \neq 0$, it is possible to rewrite the structural equation as

$$\tilde{y}_1 = \tilde{\beta} y_2 + u,$$

where $\tilde{y}_1 = y_1 - \beta_0 y_2$, and $\tilde{\beta} = (\beta - \beta_0)$, so that testing $\beta = \beta_0$ is equivalent to testing $\tilde{\beta} = 0$.

[11] An optimality result is available in the exactly identified case (when $k = 1$). As shown in Moreira (2003), the Anderson-Rubin test (Anderson and Rubin (1949)) is uniformly most powerful unbiased when $k = 1$.

[12] For an exhaustive treatment, we refer to AMS, Chamberlain (2001), and Chamberlain and Imbens (2004). In particular, Chamberlain (2001) provides a minimax justification of the invariance restriction.

the testing problem remains nonstandard even if attention is restricted to invariant testing procedures. In fact, AMS's Corollary 2 establishes that the functional form of a point optimal invariant similar test depends on the specific alternative. As a consequence, no UMP invariant similar test exists.

In this paper, we focus on tests that are similar conditional on the first stage F statistic. As discussed in detail in Section 5, the decision to condition on the first stage F statistic is motivated by a generalization of the conditionality principle and by the fact that the first stage F statistic is a specific ancillary for β, in the sense of Basu (1977). In light of the choice to condition on the first stage F statistic, it is natural to propose a characterization of the maximal invariant which lays emphasis on the specific ancillary.

Let G be equal to (G_{AR}, G_C, G_F) and be defined as a one-to-one transformation of AMS's maximal invariant, with

$$\begin{aligned} G_F &= \omega_{22}^{-1} y_2' N_Z y_2 \\ G_C &= \omega_{22.1}^{-1} \left(y_2 - \omega_{11}^{-1}\omega_{21} y_1\right)' N_Z \left(y_2 - \omega_{11}^{-1}\omega_{21} y_1\right) \\ G_{AR} &= \omega_{11}^{-1} y_1' N_Z y_1 \ , \end{aligned} \tag{5}$$

where $N_Z = Z\left(Z'Z\right)^{-1} Z'$ is the projection matrix onto the column space of the $n \times k$ matrix of instruments Z, and $\omega_{22.1} = \omega_{22} - \omega_{11}^{-1}\omega_{21}^2$ is the conditional variance of v_{i2} given v_{i1}. Lemma 2.1, a corollary of Lemma 3 of AMS, provides a characterization of the distribution of the random vector G.

Lemma 2.1 *The density of G is given by*

$$\begin{aligned} f_G\left(g_{AR}, g_C, g_F; \beta, \lambda_F\right) &= K \exp\left(-\frac{1}{2}\left(\lambda_{AR} + \lambda_C\right)\right) \\ &\times \left[g_{AR} g_C - \left(\frac{\left[g_F - \rho^2 g_{AR} - \left(1-\rho^2\right) g_C\right]}{2\rho\sqrt{1-\rho^2}}\right)^2\right]^{(k-3)/2} \exp\left\{-\frac{1}{2}\left(g_{AR} + g_C\right)\right\} \\ &\times {}_0\tilde{F}_1\left(; \frac{k}{2}; \frac{1}{4}\left[\lambda_{AR} g_{AR} + \lambda_C g_C - 2\delta\left(\beta\right)\sqrt{\lambda_{AR}\lambda_C}\frac{\left[g_F - \rho^2 g_{AR} - \left(1-\rho^2\right) g_C\right]}{2\rho\sqrt{1-\rho^2}}\right]\right), \end{aligned} \tag{6}$$

where $K^{-1} = 2^{k+1}\sqrt{\pi}\Gamma\left(\frac{k-1}{2}\right)|\rho|\sqrt{1-\rho^2}$, ${}_0\tilde{F}_1\left(; \cdot; \cdot\right)$ *denotes the regularized (confluent) hypergeometric function,* $\rho = \omega_{12}/\sqrt{\omega_{11}\omega_{22}}$, $\delta\left(\beta\right) = sign\left[\beta\left(1 - \omega_{11}^{-1}\omega_{21}\beta\right)\right]$, *and*

$$\begin{aligned} \lambda_C &= \lambda_C\left(\beta, \lambda_F\right) = \omega_{22.1}^{-1}\omega_{22}\lambda_F\left(1 - \omega_{11}^{-1}\omega_{21}\beta\right)^2 , \\ \lambda_{AR} &= \lambda_{AR}\left(\beta, \lambda_F\right) = \omega_{11}^{-1}\omega_{22}\lambda_F\beta^2. \end{aligned} \tag{7}$$

The statistics G_{AR}, G_C, and G_F are quadratic forms centered around the projection matrix N_Z, and are distributed noncentral χ_k^2, with noncentrality parameters $\lambda_{AR}\left(\beta, \lambda_F\right)$, $\lambda_C\left(\beta, \lambda_F\right)$ and λ_F, respectively. In particular, the distribution of G_F does not depend on β so that, by definition, G_F is a specific ancillary for β.

Conditioning on G_F successfully reduces the dimension of the data, delivering a problem characterized by a two-dimensional statistic and a two-dimensional parameter. Nonetheless inference on β remains nontrivial since the probability model for (G_{AR}, G_C) given G_F is not a member of the exponential family. In consequence, no standard results for the exponential family can be invoked.

Remark 2.1

The statistics G_{AR}, G_C, and G_F all have straightforward interpretations. Specifically, G_{AR} corresponds to the Anderson-Rubin (1949) statistic when the covariance matrix of the error terms is known; G_C is a complete and sufficient statistic for the concentration parameter under the null hypothesis (AMS); G_F is proportional to the first stage F statistic to test the null that $\Pi = 0$ in (2),[13] typically considered a measure of the strength of the instruments (Stock, Yogo, and Wright (2002)).

3 Point Optimal Tests

For any testing problem, an invariant test can be represented by a $[0,1]$-valued function $\phi(\cdot)$ such that the null hypothesis is rejected with probability $\phi(g)$ when $G = g$. The probability of rejecting H_0 is given by $E_{\beta,\lambda_F}\phi(G)$, where the subscripts denote the probability distribution (indexed by β and λ_F) with respect to which expectation is taken.

Making use of this notation, we formalize the definition of a conditionally similar test. An invariant size α test is said to be conditionally similar or, equivalently, to have Neyman structure with respect to G_F, if[14]

$$E_{0,\lambda_F}\left[\phi(G)\,|G_F\right] = \alpha \qquad \forall \lambda_F > 0. \tag{8}$$

The objective is to derive an (optimal) conditionally similar test for the following two-sided testing problem:

$$H_0 : \beta = 0 \qquad vs. \qquad H_1 : \beta \neq 0. \tag{9}$$

Finding a UMP test among conditionally similar tests corresponds to maximizing the conditional power function $E_{\beta,\lambda_F}\left[\phi(G)\,|G_F\right]$, uniformly in λ_F and G_F, subject to (8), or equivalently to maximizing the unconditional power $E_{\beta,\lambda_F}\left[\phi(G)\right]$, uniformly in λ_F, subject to the same conditional similarity restriction.

Theorem 3.1 focuses on testing problems of the form

$$H_0 : \beta = 0 \qquad vs. \qquad H_1 : \beta = \beta_1, \tag{10}$$

where β_1 is an arbitrary constant. A UMP conditionally similar test for (9) exists if and only if a UMP conditionally similar test exists for any testing problem of the form (10) and the functional form of the (point) optimal test does not depend on the specific value hypothesized under H_1. Theorem 3.1 provides a characterization of the optimal test for the testing problem in (10).

[13] G_F is equal to the first stage F statistic times the number of instruments. Hansen, Hausman, and Newey (2004) argue that G_F provides a better measure of the strength of the instruments than does the first stage F statistic.

[14] Throughout the paper, statements involving conditional expectations are assumed to hold almost surely.

Theorem 3.1 *If $\beta_1 \neq \sqrt{\omega_{11}/\omega_{22}}\rho$, then the UMP conditionally similar (invariant) level α test for the testing problem in (10) is given by*

$$\phi_\alpha^*(g;\rho,\beta_1) = \begin{cases} 1\left[g_{AR} > c_\alpha\left(g_C, g_F;\rho,\beta_1\right)\right] & \text{if} \quad \beta_1\left(\beta_1 - \sqrt{\frac{\omega_{11}}{\omega_{22}}}\rho\right) > 0 \\ 1\left[g_{AR} < c_\alpha\left(g_C, g_F;\rho,\beta_1\right)\right] & \text{if} \quad \beta_1\left(\beta_1 - \sqrt{\frac{\omega_{11}}{\omega_{22}}}\rho\right) < 0, \end{cases} \tag{11}$$

where $1[\cdot]$ is the indicator function and $c_\alpha(\cdot)$ is defined such that

$$E_{0,\lambda_F}\left[\phi_\alpha^*(G;\rho,\beta_1)\,|G_C, G_F\right] = \alpha \qquad \forall\lambda_F > 0. \tag{12}$$

If $\beta_1 = \sqrt{\omega_{11}/\omega_{22}}\rho$, then any conditionally similar test is optimal.

The optimal test in Theorem 3.1 has Neyman structure with respect to (G_C, G_F) (see equation (12)). It follows from Lemma A.1 of Appendix A and Lehmann (1997, Theorem 4.2) that having Neyman structure with respect to G_F is equivalent to having Neyman structure with respect to the pair (G_C, G_F), implying that any conditionally similar test has Neyman structure with respect to (G_C, G_F).

Under the null, the distribution of G_{AR} conditional on (G_C, G_F) is independent of the nuisance parameter since (G_C, G_F) is sufficient for the λ_F. As a result, the test function $\phi_\alpha^*(\cdot;\rho,\beta_1)$ is such that the critical value $c_\alpha(G_C, G_F;\rho,\beta_1)$ defined in equation (12) does not depend on λ_F.

The test in Theorem 3.1 can be thought of as an infeasible conditional Anderson-Rubin test that always rejects for large values of the G_{AR} except when β_1 is such that $\beta_1\left(\beta_1 - \sqrt{\omega_{11}/\omega_{22}}\rho\right)$ is negative. It is not uncommon for the optimal rejection rule to depend on whether the alternative lies to the left or to the right of the null, i.e. whether β_1 is greater or less than zero (as, for instance, in the case of the one-sided optimal test for location models). It is instead surprising that the optimal rejection rule changes for values of the alternative hypothesis on the same side of the null, i.e. for β_1 greater or less than $\sqrt{\omega_{11}/\omega_{22}}\rho$. (When $\beta = \sqrt{\omega_{11}/\omega_{22}}\rho$, the conditional density of G_{AR} given (G_C, G_F) is equal to the null conditional density, delivering the conclusion that any conditionally similar test is optimal.)

Theorem 3.1 provides an analytic description of the efficiency bound for conditionally similar invariant level α tests and allows us to conclude that a conditionally similar UMP test does not exist for the testing problem in (9). However, the functional form of the point optimal test is suggestive of a feasible test that meets the efficiency bound over a large portion of the parameter space. Indeed, let $\phi_\alpha^{CAR}(g;\rho)$ be the test function associated with the test that always rejects for large values of the Anderson-Rubin test statistic, i.e.

$$\phi_\alpha^{CAR}(g;\rho) = 1\left[g_{AR} > c_\alpha^{CAR}(g_C, g_F;\rho)\right],$$

where $c_\alpha^{CAR}(\cdot)$ is defined as in (12). By construction the conditional power of $\phi_\alpha^{CAR}(\cdot;\rho)$ coincides with the efficiency bound for conditionally similar tests as long as $\beta_1\left(\beta_1 - \sqrt{\omega_{11}/\omega_{22}}\rho\right)$ is positive, while differing elsewhere. It is not possible to determine analytically the magnitude of the shortcoming of the conditional AR test. Quantifying the discrepancy between the power envelope and the power curve of $\phi_\alpha^{CAR}(\cdot;\rho)$ is a numerical question that is explored in detail in Section 7.

The following corollary is a restatement of Theorem 3.1. The restatement has the advantage of not involving conditional distributions, a property that will prove useful in Section 6.2.2. (The equivalence between the two statements follows from Bierens (1982) and Lemma A.2 of Appendix A.)

Corollary 3.1 *Suppose the test function $\phi(\cdot)$ satisfies*

$$E_{0,\lambda_F}\left[(\phi(G)-\alpha)\,h(G_F)\right]=0 \qquad \forall\lambda_F>0, h\in C_b^+(\mathbb{R}_{++}), \tag{13}$$

where $C_b^+(\mathbb{R}_{++})$ denotes the set of nonnegative continuous and bounded real valued functions on $\mathbb{R}_{++}$. Then

$$E_{\beta,\lambda_F}\left[\{\phi(G)-\phi_\alpha^*(G;\rho,\beta)\}\cdot h(G_F)\right]\le 0 \qquad \forall\beta\in\mathbb{R}, \lambda_F>0, h\in C_b^+(\mathbb{R}_{++}).$$

Remark 3.1

The cut-off rule for the point optimal test can be re-expressed in terms of alternative parameters that have previously appeared in the simultaneous equations literature.

Define the standardized structural coefficient as $\eta=|\Omega|^{-1/2}\omega_{22}(\beta_1-\omega_{12}/\omega_{22})$ (Anderson, Kunitomo, and Sawa (1982) and Hahn, Hausman, and Kuersteiner (2004)).[15] Then

$$\beta_1\left(\beta_1-\sqrt{\tfrac{\omega_{11}}{\omega_{22}}}\rho\right)=\beta_1\left(\beta_1-\tfrac{\omega_{12}}{\omega_{22}}\right)\lesseqgtr 0 \quad \text{iff} \quad \beta_1\eta\lesseqgtr 0.$$

Alternatively, let $\rho_S=\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ be the structural correlation coefficient, i.e. the correlation coefficient computed from the covariance matrix Σ of the structural error terms, as defined in (3). Then

$$\beta_1\left(\beta_1-\sqrt{\tfrac{\omega_{11}}{\omega_{22}}}\rho\right)=\beta_1\left(-\tfrac{\sigma_{12}}{\sigma_{22}}\right)\lesseqgtr 0 \quad \text{iff} \quad \beta_1\rho_S\gtreqless 0.$$

Following Bound, Jaeger, and Baker (1995) and Hahn and Hausman (2002), define the finite sample bias of the two-stage estimate for β as $bias_{IV}=\lambda_F^{-1}(k-2)\,\sigma_{12}\sigma_{22}$. Then

$$\beta_1\left(\beta_1-\sqrt{\tfrac{\omega_{11}}{\omega_{22}}}\rho\right)=\beta_1\left(-\tfrac{\sigma_{12}}{\sigma_{22}}\right)\lesseqgtr 0 \quad \text{iff} \quad \beta_1 bias_{IV}\gtreqless 0.$$

Finally, AMS identify $\beta_{AR}=\omega_{11}/\omega_{12}$ as a critical value of β in their discussion of the properties of the point optimal similar tests. Specifically, AMS point out how the point optimal test changes dramatically according to position of the value hypothesized under the alternative relative to β_{AR}.[16] It turns out that β_{AR} is also related to the cut off point for the point optimal derived in Theorem 3.1.

[15] η measures the difference between β and the regression coefficient of v_1 on v_2.

[16] AMS also explain the nonmonotonicity of the LM test (Kleibergen (2002) and Moreira (2001)) making use of β_{AR}. In particular, the power function of the LM test exhibits unusual behavior as β changes from values less than β_{AR} to values greater than β_{AR}.

In particular, β_{AR} is proportional to the inverse of $\sqrt{\omega_{11}/\omega_{22}}\rho$:

$$\beta_1\left(\beta_1-\sqrt{\tfrac{\omega_{11}}{\omega_{22}}}\rho\right)=\beta_1\left(\beta_1-\tfrac{\omega_{12}}{\omega_{22}}\right)\lesseqgtr 0 \quad \text{iff} \quad \beta_1\left(\beta_1-\tfrac{\omega_{11}}{\omega_{22}}\tfrac{1}{\beta_{AR}}\right)\lesseqgtr 0.$$

Remark 3.2

In the context of the model of Section 2.1, Moreira (2003) develops a general method for constructing similar tests based on the conditional distribution of nonpivotal statistics given G_C.[17] In Moreira (2003), the relevant measure of performance is unconditional power and conditioning is used as a tool to characterize similar tests. In this paper, the relevant measure of performance is assumed to be conditional power (i.e., power conditional on G_F), in which case the natural similarity restriction is the conditionally similarity restriction (8). Even if the relevant criterion is unconditional power, then it is possible to view the conditional similarity restriction as an alternative way of developing tests with correct size since, by the law of iterated expectations, any conditionally similar test is also similar. (However, imposition of the conditional similarity restriction is likely to induce a loss in unconditional power.)

4 Optimal Conditionally Unbiased Tests

In the context of exponential families, optimality results for two-sided testing problems are obtained by requiring the testing procedure to be unbiased (Lehmann (1997, Chapter 4)). We impose the same restriction in order to derive a UMP conditionally unbiased test for the two-sided testing problem in (9).

Making use of the notation introduced in Section 3, an invariant size α test, with test function $\phi(\cdot)$, is conditionally unbiased if and only if its power function is such that

$$\begin{aligned} &E_{0,\lambda_F}\left[\phi(G)\,|G_F\right]\leq\alpha &&\forall\lambda_F>0,\\ &E_{\beta,\lambda_F}\left[\phi(G)\,|G_F\right]\geq\alpha &&\forall\beta\neq 0,\lambda_F>0. \end{aligned} \tag{14}$$

It turns out (see proof of Lemma 4.1, in Appendix B) that the conditional power functions of tests based on G are continuous and differentiable in β, and that derivatives with respect to β can be computed by differentiating under the integral sign. These properties enable us to characterize a set of necessary conditions for an invariant test to be conditionally unbiased. In particular, continuity of the power function implies that any conditionally unbiased test is also conditionally similar (i.e., satisfies (8)). Furthermore, it follows from the definition in (14) and from the formulation of the testing problem in (9) that the power function for a conditionally unbiased test must have a minimum at $\beta=0$. The

[17] Building on Moreira (2003), AMS show that an invariant test is similar if and only if it has Neyman structure with respect to G_C.

differentiability of the power function with respect to β implies

$$\left.\frac{\partial}{\partial\beta}E_{\beta,\lambda_F}\left[\phi\left(G\right)|G_F\right]\right|_{\beta=0}=0 \qquad \forall\lambda_F>0, \tag{15}$$

Relying on (8) and (15), Lemma 4.1 provides a set of necessary conditions for a test to be conditionally unbiased.

Lemma 4.1 *An invariant size α test, with test function $\phi(\cdot)$, is conditionally unbiased only if*

$$\begin{aligned} E_{0,\lambda_F}\left[\phi\left(G\right)|G_C,G_F\right] &= \alpha && \forall\lambda_F>0, && (16)\\ E_{0,\lambda_F}\left[\phi\left(G\right)G_{AR}|G_C,G_F\right] &= \alpha E_{0,\lambda_F}\left[G_{AR}|G_C,G_F\right] && \forall\lambda_F>0. && (17)\end{aligned}$$

Invoking Lemma 4.1 and the generalized Neyman-Pearson Lemma (Lehmann (1997, Theorem 3.5)), Theorem 4.1 characterizes the optimal conditionally unbiased test for the two-sided testing problem.

Theorem 4.1 *The UMP (invariant) conditionally unbiased level α test for the testing problem in (10) is given by*

$$\phi_\alpha^{**}\left(G;\rho\right)=1\left[g_{AR}<c_{1,\alpha}\left(g_C,g_F;\rho\right)\right]+1\left[g_{AR}>c_{2,\alpha}\left(g_C,g_F;\rho\right)\right], \tag{18}$$

where $c_{1,\alpha}(\cdot,\cdot,\rho)$ and $c_{2,\alpha}(\cdot,\cdot,\rho)$ are defined such that (16) and (17) are satisfied.

As in the exponential family framework, continuity and differentiability of the conditional power function of any test based on G play a key role in the proofs of Lemma 4.1 and Theorem 4.1. Furthermore, in order to conclude that the rejection rule for the optimal conditionally unbiased test is two-sided (as in the exponential family), we exploit the fact that the regularized confluent hypergeometric function ${}_0\tilde{F}_1\left(;\frac{k}{2};\cdot\right)$ is strictly convex (as is the exponential function) and that the argument of ${}_0\tilde{F}_1\left(;\frac{k}{2};\cdot\right)$ is an affine transformation of G_{AR}.

Although the probability models for G and for (G_{AR},G_C) given G_F are not members of the exponential family, the results in Lemma 4.1 and in Theorem 4.1 parallel those derived in the standard exponential family framework (e.g., Lehmann (1997, Chapter 4.2)). The analogy between the results of this section and those in the exponential family framework is somewhat surprising in light of the results of the previous section. In particular, the result of Theorem 3.1 cannot be traced or interpreted by means of an analogous result in the exponential case.

As in the previous section, the optimality of the testing procedure based on $\phi_\alpha^{**}(G;\rho)$ can be re-expressed by means of a statement that does not involve conditional distributions.

Corollary 4.1 *Suppose the test function $\phi(\cdot)$ satisfies*

$$\begin{aligned} E_{0,\lambda_F}\left[\left(\phi\left(G\right)-\alpha\right)\cdot h\left(G_F\right)\right] &= 0 && \forall\lambda_F>0, h\in C_b^+\left(\mathbb{R}_{++}\right),\\ E_{0,\lambda_F}\left[\left(\phi_\alpha^{**}\left(G;\rho\right)-\alpha\right)G_{AR}\cdot h\left(G_F\right)\right] &= 0 && \forall\lambda_F>0, h\in C_b^+\left(\mathbb{R}_{++}\right).\end{aligned}$$

Then

$$E_{\beta,\lambda_F}\left[\left\{\phi\left(G\right)-\phi_{\alpha}^{**}\left(G;\rho\right)\right\}\cdot h\left(G_F\right)\right]\leq 0 \qquad \forall\beta\in\mathbb{R},\lambda_F>0,h\in C_b^{+}\left(\mathbb{R}_{++}\right).$$

The "unconditional" formulations of the (finite sample) optimality results in Corollaries 3.1 and 4.1 will form the basis of our development of asymptotic counterparts of Theorems 3.1 and 4.1. The corollaries simplify the characterization of the limiting behavior of the testing procedures formulated in Sections 3 and 4 by enabling us to avoid the complexities associated with limits of conditional distributions (e.g., Sweeting (1989)).

5 Conditional vs Unconditional Inference

This is not the first paper to emphasize the importance of the first-stage F statistic and/or to suggest a conditional approach in the context of the simultaneous equations model. By arguing in favor of pretesting, Hansen, Hausman, and Newey (2004), Staiger and Stock (1997), Stock and Yogo (2002), Wang and Zivot (1998), and Zivot, Startz, and Nelson (1998) are implicitly suggesting conditional two-step procedures that condition discretely on the first-stage F statistic. More recently, Forchini and Hillier (2003) insist that reported precision of parameter estimates, such as standard errors and mean squared errors, should be conditional on the observed value of an identification test statistic.[18] Starting from a generalization of the notion of ancillary statistics and of the conditionality principle, we come to an analogous conclusion in the context of inference, and we therefore focus on similar tests conditional on the first stage F statistic.

In small samples, conditional procedures can be less efficient than their unconditional counterparts, however in many cases, the (efficiency) difference tends to disappear as the sample size becomes large (see Lehmann and Scholz (1992)). For instance, in the classical linear regression framework with random regressors, asymptotically there is no difference between carrying out inference conditional on the regressors or unconditionally.

In those instances in which the discrepancy between conditional and unconditional power curves does not vanish in the limit, several authors (Severini (1995), Cox (1988), and Lehmann (1997)) have pointed out that the loss of unconditional power may not be a proper measure of comparison if conditioning makes probability calculations more relevant to the data at hand.

In the case of the simultaneous equations model, adopting the conditional or unconditional point of view yields different conclusions, as is confirmed by the numerical results in Section 7. Hence, it is particularly important to motivate the choice between the conditional and unconditional approaches and to establish which framework is more pertinent to the particular application.

Unfortunately, when and whether conditional inference is appropriate is a rather controversial subject (see Reid (1995), Lehmann and Scholz (1992), Cox and Hinkley (1974)). The controversy originates in part from the fact that whether conditioning is opportune depends on the repeated

[18] The analytical and numerical work of Forchini and Hillier (2003) uses the first stage F statistic as the identification test statistic. Alternative identification test statistics have been proposed by Forchini and Hillier (2004).

sampling underlying the model of interest (Lehmann (1997, pp. 539-541), Lehmann and Scholz (1992), and Kalbfleisch (1975)). That is, conditioning cannot be justified solely on the basis of mathematical properties of the parametric model under consideration, but depends on the context of the experiment underlying the model.

5.1 A Case for Conditioning

In this section, we argue that in the case of the endogenous regressors model, the set of problems for which the proper probability calculations involve the conditional distribution of (G_{AR}, G_C) given G_F contains commonly used statistical models. That is, in certain economic applications, it is desirable to condition on the design of the experiment that has actually been run.

In discussing this point, we first describe the classical Cox example (Cox (1958)), since it represents a situation in which conditioning on ancillaries is noncontroversial (Cox and Hinkley (1974)). In particular, we emphasize those features of the example that allow us to conclude that conditioning is legitimate. Then, by means of examples, we argue that in some cases, the experiment underlying the IV problem has a similar structure to that of a Cox-type example, and therefore the conditional point of view is the more relevant one.[19]

Example 5.1 (Cox (1958)) *An experiment is conducted as follows: a fair coin is tossed and, according to the outcome of the toss (randomization mechanism), a measuring device is selected and employed to perform another experiment, such as measuring the length of a table. The two measuring devices differ only in terms of their precision. The random variables associated with each measuring device are characterized by a mean equal to the true length of the table, denoted by μ, and by variances equal to 1 and 100 for the precise and imprecise instruments, respectively.*

Suppose we are interested in conducting inference on μ, and that the imprecise measuring device was selected by the coin toss (sub-experiment 1), and the second part of the experiment is performed. The actual measurement, denoted by x, is an estimate of the true length of the table (μ).

Conditional inference concerning μ holds fixed the measuring device actually used, while the information that some other measuring device might have produced the data is disregarded, yielding a standard error for the estimate of $\sqrt{100}$. However, if we choose not to condition on the realization of the coin toss, the standard error would be $\sqrt{50.5}$. Adopting the unconditional point of view leads to a dramatic reduction of the standard error. This reduction is arguably artificial, since it results from taking into consideration the fact that the measurement could have been produced by a more accurate instrument, which was not employed in the experiment.

The Cox type example is considered to be a very compelling case in which inference about μ should be conditional on the realization of the ancillary statistic, namely the outcome of the randomization device (see Cox (1958), Cox and Hinkley (1974), Lehmann and Scholz (1992)). The latter can be regarded as an indicator of the part of the total sample space relevant to the problem at hand.

[19] In Appendix C, we make the parallel between the Cox-type example discussed in this Section and the IV model explicit.

In addition to the fact that the coin toss is an ancillary statistic for μ,[20] what makes the idea of conditioning in the Cox example so appealing is twofold:

1. The (sub) experiments occur sequentially: we first observe the coin toss and conditional on its realization, we perform the measurement of the table. The data from the two sub-experiments can be represented by a bivariate random vector, whose joint density can be factorized into the product of the conditional density of the measurement of the table given the outcome of the coin toss times the marginal probability distribution for the coin toss. In light of the sequential nature of the experiment, this factorization is informative of the temporal dynamics of the experiment.

2. Finally, knowledge of the probabilities of observing heads or tails does not improve our knowledge about the true length of the table, since the range of possible values of μ is unrestricted. Conversely, knowing the true length of the table does not inform us about the parameters governing the coin toss.

Now, consider the model in Section (2.1). Sufficiency and invariance enabled us to reduce the data to the tridimensional vector (G_{AR}, G_C, G_F), where G_F is a specific ancillary for β. Mathematically, it is always possible to represent the joint density of (G_{AR}, G_C, G_F) as a product of the conditional density of (G_{AR}, G_C) given G_F and the marginal of G_F. In the spirit of Kalbfleisch (1975) and Basu (1964), we argue that conditioning on G_F is desirable when this factorization provides a more accurate description of how the data is actually generated, and is not a simple mathematical identity.[21]

To establish whether conditioning reflects the structure of the underlying experiment, we need to analyze the specific context of the IV model. The mathematical properties of the model are not sufficient to justify the choice to condition on the first stage F. To illustrate this point, consider the following two examples. In both cases, the data are assumed to originate from the model in Section 2.1, but in the first one, conditioning does not seem justifiable, while in the second it does. In discussing each example we emphasize when and whether the analogy with the Cox example is applicable.

Example 5.2 (Demand for Cigarettes) *Consider the model in Section (2.1) and assume that y_1 measures cigarette consumption, while y_2 measures the price of cigarettes (including all taxes in the state); also suppose that we have at least two exogenous instruments (e.g. the portion of the tax on cigarettes arising from general sales taxes and the amount of cigarette-specific taxes).*[22]

[20] That is, the probability of obtaining heads is completely known. It is possible to formulate Cox type examples associated with generalizations of the definitions of ancillary statistics, the basic idea being that the argument for conditioning remains valid even if the coin is biased (with unknown bias) as long as the bias is not a function of μ. In particular, the notion of ancillarity can be adapted to those cases in which nuisance parameters are present (see Basu (1977) and Cox and Hinkley (1974, pages 32-35)).

[21] Basu (1964) emphasizes the importance of distinguishing between performable and nonperfomable experiments. A similar argument is made by Kalbfleisch (1975), who introduces the (undefined) notions of experimental and mathematical mixtures in an attempt to make progress on a formalization of the distinction emphasized by Basu (1964).

[22] For simplicity, suppose we are in a closed economy, and that the market for cigarettes is competitive and always clears. In addition assume that the data available to us is a time series on equilibrium prices and quantities.

The mixture model (Lehmann (1997) and Kalbfleisch (1982)) corresponding to the cigarette example could be described as follows. We first observe the realization of prices, determined according to suppliers' price setting rule as a function of cigarette taxes (reduced form equation). For given prices, consumers choose the optimal consumption of cigarettes (structural equation).

However the reduced form equation is not invariant to realizations of the demand, since firms' optimal pricing policies take the demand curve they face as given. Therefore, the reduced form error terms are by construction functions of the structural equation's error terms. Equilibrium prices and quantities are determined simultaneously and there are continuous feedbacks between reduced form and structural equations. This feature of the data (feedbacks between the two equations) would be ignored by conditioning on the first stage F statistic.

Example 5.3 (Return to Schooling) *Consider the model in Section (2.1) and suppose that y_1 measures earnings, while y_2 measures education; the agent has completed the desired amount of schooling and we assume that endogeneity arises exclusively because of an omitted variable, students' ability, and that we have at least two exogenous instruments.*

We argue that, in the return to schooling example, the choice of the optimal amount of education (reduced form) is analogous to the coin toss sub-experiment; while the labor market's valuation of individual productivity (earnings), which is a function of education and unobservable ability, corresponds to the measurement of the table in the Cox experiment. Unlike in Example 5.1, the two sub-experiments are statistically "related" through the structure of the error terms. However, the source of correlation between disturbances results exclusively from ability being an omitted variable.[23]

Under the assumptions made, the optimal amount of schooling is determined before observing earnings realizations and the preference parameters that govern this choice are invariant to parameters and realizations of the earning equation. The parameters of the reduced form equations do not depend on β, nor is the structural coefficient a function of Π. For instance, variation in the correlation between mother's education and the chosen level of schooling does not affect the way firms price education, since the firm is only interested in the amount of education achieved. Similarly, a change in how the labor market values education (β) does not alter the way mother's education affects the optimal schooling decision (Π). In addition, the decision regarding how much to invest in schooling occurs prior to the realization of earnings, making the sequential structure of the Cox example a feature of the return to schooling example.

Conditioning in the return to schooling example reflects the nature of the experiment underlying the IV model. Therefore probability calculations involving the conditional distribution of (G_{AR}, G_C) given G_F provide a better description of the data at hand.

In Appendix C, we show that when there exists a control variable (Blundell and Powell (2004) and Newey, Powell, and Vella (1999)), as in the return to schooling example, and when knowledge of

[23] We are ruling out Becker-type models of human capital formation (Becker (1967)) in which expectations on future earnings are determinants of the observed amount of schooling (for a review of the return to schooling literature, see Card (2001)). However it could be possible to allow some form of expectation to enter in the schooling decision as long as the agent does not reoptimize dynamically as he acquires information about realized earnings or changes in β.

the parameters governing the reduced form do not inform us about the parameter of the structural equation (and vice versa), the data underlying the IV model can be thought of as being generated by a Cox-type experiment. This experiment can be described as follows: two tosses of a coin (with unknown probabilities of observing heads or tails) determine which experiment to run, but only one toss is observed and corresponds to the first stage F statistic (G_F). Even if we do not observe the realizations of both coin tosses, the outcome of one coin toss still contains information that can be used to rule out states of the world that are not relevant for the experiment actually conducted.

6 Extensions

This section considers two extensions of the Gaussian model of Section 2.1. Section 6.1 augments the Gaussian model by introducing exogenous regressors, while Section 6.2 relaxes the assumptions of Gaussian errors and known covariance matrix.

6.1 Exogenous Regressors

Consider the model

$$y_1 = \beta y_2 + X\tilde{\gamma} + u, \tag{19}$$

$$y_2 = Z\Pi + X\theta + v_2, \tag{20}$$

where y_1 and y_2 are n-dimensional vectors of endogenous variables, X is an $n \times k_1$ matrix of exogenous regressors, Z is an $n \times k_2$ matrix of instruments, the matrix $[X : Z]$ has full column rank $(k_1 + k_2)$, $k_2 > 1$, and we assume, without loss of generality, that $X'Z = 0$.[24]

The reduced form equations are

$$\begin{aligned} y_1 &= Z\Pi\beta + X\gamma + v_1, \\ y_2 &= Z\Pi + X\theta + v_2, \end{aligned}$$

where $\gamma = \tilde{\gamma} + \theta\beta$ and $v_1 = u + \beta v_2$.

Under the same assumptions on the error terms as in Section 2.1, there is no loss of attainable power by restricting our attention to tests which depend on Y only through the minimal sufficient statistic $[X : Z]'Y$.

As pointed out by Moreira (2003), one and two sided testing problems concerning β are invariant under translations of the form $Y + XF$, for any conformable matrix F. $Z'Y$ is a maximal invariant

[24] Whenever $X'Z \neq 0$, it is possible to rewrite the reduced form equations in terms of the orthogonal pair $\left[X : Z^{\perp}\right]$:

$$\begin{aligned} y_1 &= Z^{\perp}\Pi\beta + X\bar{\gamma} + v_1 \\ y_2 &= Z^{\perp}\Pi + X\bar{\theta} + v_2 \ , \end{aligned}$$

where $Z^{\perp} = M_X Z$, $\bar{\theta} = \theta + N_X Z\Pi$, $\bar{\gamma} = \gamma + \bar{\theta}\beta$, $N_X = X\left(X'X\right)^{-1}X'$and $M_X = I_{k_1} - N_X$ are the projection matrices onto the column and the null spaces of X respectively.

under the same group of transformations and its distribution does not depend on (θ, γ).[25] Therefore, sufficiency, invariance, and Lehmann (1997, Theorem 6.3) enable us to eliminate the $2k_1$ nuisance parameters (θ, γ), yielding the same framework we considered in Sections 2.2 and 3. As a consequence, Theorems 3.1 and 4.1 apply to the model described by equations (19) and (20).

6.2 Asymptotics

6.2.1 Weak Instruments Asymptotics

The models in Sections 2.1 and 6.1 can further be generalized by relaxing the assumptions of Gaussian errors and known covariance matrix. Conventional asymptotics (in which Π is held fixed) provide poor approximations when the instruments can be arbitrarily weak, whereas Staiger and Stock (1997) type of asymptotics (in which $\Pi = O(1/\sqrt{n})$) has been found to provide good approximations in weak instruments settings. In recognition of the empirical relevance of the weak instruments problem, we therefore rely on Staiger and Stock (1997) type of asymptotics when characterizing the limiting conditional power of (feasible versions of) the tests derived in Sections 3 and 4.

Let $\hat{G} = \left(\hat{G}_{AR}, \hat{G}_C, \hat{G}_F\right)$, where[26]

$$\begin{aligned} \hat{G}_F &= \hat{\omega}_{22}^{-1} y_2' N_Z y_2, \\ \hat{G}_C &= \hat{\omega}_{22.1}^{-1} \left(y_2 - \hat{\omega}_{11}^{-1}\hat{\omega}_{21} y_1\right)' N_Z \left(y_2 - \hat{\omega}_{11}^{-1}\hat{\omega}_{21} y_1\right), \\ \hat{G}_{AR} &= \hat{\omega}_{11}^{-1} y_1' N_Z y_1, \end{aligned}$$

and

$$\hat{\Omega} = \begin{bmatrix} \hat{\omega}_{11} & \hat{\omega}_{12} \\ \hat{\omega}_{21} & \hat{\omega}_{22} \end{bmatrix} = (Y - N_Z Y)' (Y - N_Z Y)/n.$$

Following Staiger and Stock (1997), we assume:
$(A1)$ $\Pi = D/\sqrt{n}$, where D is a nonrandom $k \times 1$ vector.
$(A2)$ $V'V/n \xrightarrow{p} \Omega$, where $V = [v_1 : v_2]$ and Ω is a 2×2 positive definite matrix.
$(A3)$ $Z'Z/n \xrightarrow{p} Q_{ZZ}$, where Q_{ZZ} is a positive definite matrix.
$(A4)$ $Z'V/\sqrt{n} \xrightarrow{d} N(0, \Omega \otimes Q_{ZZ})$.

An immediate implication of the above set of assumptions is

$$\lambda_F = \omega_{22}^{-1} \Pi' Z' Z \Pi = \omega_{22}^{-1} D' \frac{Z'Z}{n} D \xrightarrow{p} \lambda_{F,\infty} = \omega_{22}^{-1} D' Q_{ZZ} D.$$

As shown in Theorem 6.1, $A1 - A4$ furthermore enable us to characterize the limiting behavior of $\hat{G}$, $\phi_\alpha^*\left(\hat{G}; \hat{\rho}, \beta\right)$, and $\phi_\alpha^{**}\left(\hat{G}; \hat{\rho}\right)$, where $\hat{\rho} = \hat{\omega}_{12}/\sqrt{\hat{\omega}_{11}\hat{\omega}_{22}}$. Henceforth, let

$$\mathcal{G}(\beta, \lambda_F) = \left(\mathcal{G}_{AR}(\beta, \lambda_F), \mathcal{G}_C(\beta, \lambda_F), \mathcal{G}_F(\lambda_F)\right),$$

[25] $vec(Z'Y)$ is normally distributed with mean $vec(Z'Z\Pi(\beta : 1))$ and covariance matrix $\Omega \otimes Z'Z$.
[26] For notational simplicity, we suppress the dependence of $\hat{G}$ and $\hat{\Omega}$ on the sample size n.

be a random vector, identified by the parameters $(\beta, \lambda_F) \in \mathbb{R} \times \mathbb{R}_{++}$, whose distribution is the same as that of G in Lemma 2.1. In addition, let subscripts on "E" denote the probability distribution (indexed by β and D) with respect to which expectation is taken.

Theorem 6.1 *Under Assumptions $A1 - A4$,*

$$\hat{G} \xrightarrow{d} \mathcal{G}\left(\beta, \lambda_{F,\infty}\right), \tag{21}$$
$$\phi_\alpha^*\left(\hat{G}; \hat{\rho}, \beta\right) \xrightarrow{d} \phi_\alpha^*\left(\mathcal{G}\left(\beta, \lambda_{F,\infty}\right); \rho, \beta\right)$$
$$\phi_\alpha^{**}\left(\hat{G}; \hat{\rho}\right) \xrightarrow{d} \phi_\alpha^{**}\left(\mathcal{G}\left(\beta, \lambda_{F,\infty}\right); \rho\right)$$

In particular, for any $h \in C_b^+(\mathbb{R}_{++})$,

$$\lim_{n\to\infty} E_{\beta,D}\left[\phi_\alpha^*\left(\hat{G}; \hat{\rho}, \beta\right) h\left(\hat{G}_F\right)\right] = E\left[\phi_\alpha^*\left(\mathcal{G}\left(\beta, \lambda_{F,\infty}\right); \rho, \beta\right) h\left(\mathcal{G}_F\left(\lambda_{F,\infty}\right)\right)\right], \tag{22}$$
$$\lim_{n\to\infty} E_{\beta,D}\left[\phi_\alpha^{**}\left(\hat{G}; \hat{\rho}\right) h\left(\hat{G}_F\right)\right] = E\left[\phi_\alpha^{**}\left(\mathcal{G}\left(\beta, \lambda_{F,\infty}\right); \rho\right) h\left(\mathcal{G}_F\left(\lambda_{F,\infty}\right)\right)\right]. \tag{23}$$

Expressions (22) and (23) in Theorem 6.1 are motivated by the formulations of the optimality results in Corollaries 3.1 and 4.1 and establish that $\phi_\alpha^*\left(\hat{G}; \hat{\rho}, \beta\right)$ and $\phi_\alpha^{**}\left(\hat{G}; \hat{\rho}\right)$ have asymptotic conditional power functions equal to those of the tests derived under the assumptions of Gaussian errors and known covariance matrix.

Assumptions $A1 - A4$ do not allow us to conclude that tests based on $\phi_\alpha^*\left(\hat{G}; \hat{\rho}, \beta\right)$ and on $\phi_\alpha^{**}\left(\hat{G}; \hat{\rho}\right)$ are optimal. Only if we assume Gaussian errors does focusing exclusively on tests based on the minimal sufficient statistic $Z'Y$ not imply loss of attainable power. Therefore, in the Section 6.2.2 we augment $A1 - A4$ with a normality assumption in order to develop asymptotic counterparts of Theorem 3.1 and Theorem 4.1.

6.2.2 Asymptotic Efficiency with Gaussian Errors

We now obtain asymptotic optimality results for a sequence of experiments given by Gaussian models of the form (1)-(2) with fixed instruments, known reduced form covariance matrix, and with Π modelled local to zero as in *(A1)*. Under these assumptions, there is no efficiency loss associated with the reduction (by sufficiency) to $Z'Y$. Imposing the same invariance argument as in Section 2.2, the data are further reduced to the tridimensional vector $G = (G_{AR}, G_C, G_F)$.

In the spirit of Le Cam and Yang (2000) and van der Vaart (1998), we obtain efficiency bounds by means of the theory of limits of experiments. The derivation of these efficiency bounds is greatly facilitated by the fact that the statistical properties of the limiting experiment are isomorphic to those of the finite sample Gaussian model of Section 2.1. Notably, it follows from Theorem 6.1 that the limiting distribution of the vector G has the same properties that we exploited when deriving the optimality results for conditionally similar tests in the finite sample framework. In particular, $\mathcal{G}_F\left(\lambda_{F,\infty}\right)$ is a specific ancillary for β and $\mathcal{G}_C\left(\beta, \lambda_{F,\infty}\right)$ given $\mathcal{G}_F\left(\lambda_{F,\infty}\right)$ is complete and sufficient for the nuisance parameter $\lambda_{F,\infty}$ under the null.

As in the previous section, to derive conditional asymptotic results we reformulate conditional statements by means of unconditional moments restrictions, thus avoiding limits of conditional distributions. Using the same notation as in Section 3 and following Feigin (1986) and Jansson and Moreira (2004), a sequence of tests with asymptotic size α and test functions $\{\phi_n(\cdot)\}$ is asymptotically conditionally similar if

$$\lim_{n\to\infty} E_{0,D}\left[\left(\phi_n(G)-\alpha\right)h(G_F)\right]=0 \qquad \forall\lambda_F>0,\ h\in C_b^+(\mathbb{R}_{++}). \tag{24}$$

Similarly, a sequence of conditionally similar tests is asymptotically conditionally unbiased only if

$$\lim_{n\to\infty} E_{0,D}\left[\left(\phi_n(G)-\alpha\right)G_{AR}h(G_F)\right]=0 \qquad \forall\lambda_F>0,\ h\in C_b^+(\mathbb{R}_{++}). \tag{25}$$

Theorem 6.2 proves the asymptotic efficiency of the testing procedure derived in Theorems 3.1 and 4.1 under $(A1)$ and the following modifications of $(A2)$-$(A3)$:
$(A2')$ $V_i=[v_{1i}:v_{2i}]\backsim N(0,\Omega)$, where Ω is a known 2×2 positive definite matrix.
$(A3')$ $Z'Z/n\to Q_{ZZ}$, where Q_{ZZ} is a positive definite matrix.

Theorem 6.2 *Consider the model (1)-(2) and suppose that* $A1$, $A2'$, *and* $A3'$ *are satisfied.*
(a) If $\{\phi_n(\cdot)\}$ *satisfies (24), then*

$$\overline{\lim}_{n\to\infty}E_{\beta,D}\left[\phi_n(G)h(G_F)\right]\leq E\left[\phi_\alpha^*\left(\mathcal{G}(\beta,\lambda_{F,\infty});\rho,\beta\right)h\left(\mathcal{G}(\lambda_{F,\infty})\right)\right] \qquad h\in C_b^+(\mathbb{R}_{++}).$$

(b) If $\{\phi_n(\cdot)\}$ *satisfies (24) and (25), then*

$$\overline{\lim}_{n\to\infty}E_{\beta,D}\left[\phi_n(G)h(G_F)\right]\leq E\left[\phi_\alpha^{**}\left(\mathcal{G}(\beta,\lambda_{F,\infty});\rho\right)h\left(\mathcal{G}(\lambda_{F,\infty})\right)\right] \qquad h\in C_b^+(\mathbb{R}_{++}).$$

The assumption of normal errors permits us to develop finite sample and asymptotic efficiency bounds for conditionally similar tests and to assert that no conditionally similar (unbiased) test can improve on $\phi_\alpha^*(\cdot;\rho,\beta)$ $(\phi_\alpha^{**}(\cdot;\rho))$.

Theorem 6.2 also shows that the limiting power functions obtained under Staiger and Stock (1997) asymptotics in Theorem 6.1 are asymptotically optimal when hypotheses of uncorrelated normal homoskedastic errors and fixed instruments are met.

In light of this last remark, the normal distribution can be considered the least favorable distribution (see also Jansson and Moreira (2004)) since the maximal attainable power under the assumption of Gaussian errors and known covariance matrix can also be achieved (asymptotically) by the same testing procedures when no distributional assumption is made on the random disturbances.

Remark 6.1

Because $\hat{G}$ and $\hat{\Omega}$ are invariant under full-rank orthogonal transformations of Z, any (possibly randomized) test based on $\hat{G}$ and $\hat{\Omega}$ has power function coinciding with the power function of a (possibly randomized) test based on G. The asymptotic optimality results of this section therefore apply to sequences of tests based on $\hat{G}$ and $\hat{\Omega}$. In particular, it follows from Theorems 6.1 and 6.2 that under the assumptions of Theorem 6.2, $\phi_\alpha^*\left(\hat{G};\hat{\rho},\beta\right)$ and $\phi_\alpha^{**}\left(\hat{G};\hat{\rho}\right)$ are asymptotically optimal within the class of invariant procedures satisfying (24) and (24)-(25), respectively.

7 Monte Carlo Simulations

A Monte Carlo study is conducted to assess the performance of the conditional procedures developed in Sections 3 and 4.

The data are generated from the model in Section 2.1, maintaining the assumptions of normal errors and known covariance matrix. Without loss of generality, the hypothesized value of β under H_0 is equal to zero, and the variances ω_{11} and ω_{22} of the reduced form errors are normalized to 1. Therefore, the relevant parameters of the Monte Carlo study are k, the reduced form correlation coefficient ρ, and the concentration parameter $\lambda_F = \Pi' Z' Z \Pi$.

As in Moreira (2003) and Staiger and Stock (1997), the elements of the matrix of instruments Z are randomly drawn from a standard normal distribution and are then held fixed. Without loss of generality, we do not include exogenous regressors. Results are computed for $k = 2$, 4, 10, 50; $\rho = 0.05$, 0.2, 0.5, 0.6, 0.8, 0.95, and 0.99; and $\lambda_F/k = 0.1$ (very weak instruments), 0.5, 1, 2, 4, and 10 (strong instruments).[27]

The power plots are based on 5,000 replications, while conditional size histograms in Figures 2A and 2B are based on 500,000 replications. Monte Carlo simulations are used to compute critical values for Moreira's (2003) CLR, as suggested in Moreira (2003, Appendix C). Critical values for the conditional power envelope and for the conditionally similar test that rejects for large values of G_{AR} are obtained by numerical integration making use of Lemma A.3, which characterizes the conditional null distribution of G_{AR} given (G_C, G_F). The significance level of the tests is set equal to 5%.

To save space, we present only a representative subset of the results that have been generated.

7.1 Numerical Results

The first set of numerical results compares the power envelope for conditionally similar tests characterized in Theorem 3.1 with the power of other testing procedures such as the unconditional Anderson-Rubin test (AR), the LM test (KLM) and Moreira's (2003) CLR test (MCLR). The comparison, made in terms of unconditional power, aims to quantify the efficiency loss related to the conditional similarity restriction.

[27] The vector Π is set to be proportional to a vector of ones, with proportionality factor chosen so that λ_F/k has the desired value. Because all the testing procedures considered are invariant under full-rank orthogonal transformations of the matrix of intruments Z, it does not matter how Π is constructed.

In Figure 1A, we present the unconditional power curves for AR, KLM and MCLR and the Conditional Power Envelope (CPE) derived in Theorem 3.1.[28] If the relevant criterion to be maximized is unconditional power, then conditioning on the first stage F statistic causes a nontrivial loss of power, especially for small values of k and ρ. However, as ρ increases and for values of β greater than ρ, the CPE is not uniformly dominated by any of the other testing procedures.[29] This pattern is accentuated for larger values of k.

When $\beta = \rho$ (indicated by the solid vertical line), the CPE exhibits a curious behavior. Considering that we have normalized $\omega_{11} = \omega_{22} = 1$ and that ρ is chosen to be positive, the CPE corresponds to the power of the conditionally similar test which rejects for large values of G_{AR} if $\beta < 0$ or if $\beta > \rho$, while rejecting for small values of G_{AR} elsewhere. In light of the formulation of the power envelope and the fact that it is continuous in β, it is not surprising that power equals the size of the test (5%) when $\beta = \rho$. Simple algebra shows that when $\beta = \rho$, the density of G_{AR} conditional on (G_C, G_F) is equal to the null conditional density and therefore, at that point, CPE has power equal to the size.[30]

Although the functional form of the test that achives the CPE does not depend on the nuisance parameter, the maximum attainable power depends on the true value of λ_F. Therefore, in Figure 1B, we consider different values of λ_F/k. The shape of the power curves varies considerably as we change λ_F/k. Therefore, in order to present comparable figures, we follow AMS and plot the power curves against a rescaled alternative $\beta\sqrt{\lambda_F}$. The discrepancy between the CPE and the power curves of the other tests (AR, KLM and MCLR) shrinks for small λ_F/k, while increasing as λ_F/k increases. In particular, when the instruments are very weak and for high values of the correlation coefficient ρ, the conditional power envelope performs well, especially when β is greater than ρ.[31]

While Figures 1A and 1B illustrate that insisting on conditioning has nontrivial costs in terms of unconditional power, it does not provide information about how AR, KLM and MCLR trade conditional size for power. Drawing the analogy with some estimation problems in which large gains in efficiency can be achieved by allowing a little bit of bias, improvements in terms power could be attained without sacrificing too much in terms of conditional size. Therefore, it is of interest to investigate whether AR, KLM and MCLR, which are not conditionally similar procedures by design, are even approximately conditionally similar, or whether their gains in terms of power are obtained by "choosing" critical values as a function of G_F.

In Figures 2A and 2B, we present the histograms of rejection rates (at the null) as a function of G_F, for $k = 4$ and 10. Even for small values of ρ, it is possible to detect systematic ways in which AR,

[28] We have normalized ρ to be positive. If we were to consider negative values of ρ, all power curves would be the mirror images around 0 of those presented in Figure 1.

[29] For $\beta > \rho$, the power envelope corresponds to the power of a test which rejects for large values of AR, where the critical value is determined under the null conditional distribution of AR (conditional on G_C and G_F).

[30] In addition, it is possible to show that when $\beta = \rho$, G_F is complete and sufficient for λ_F.

[31] In light of the normalizations imposed in this section, the correlation coefficient ρ, expressed as a function of an arbitrary null $\beta = \beta_0$ and arbitrary Ω, is given by

$$\rho = \frac{\omega_{12} - \beta_0\omega_{22}}{\sqrt{\omega_{22}}\sqrt{\omega_{11} - 2\beta_0\omega_{12} + \beta_0^2\omega_{22}}}.$$

It follows from this expression that values of ρ close to unity are not unrealistic nor extreme.

KLM and MCLR choose their critical values as a function of the observed value of G_F.

We can identify at least two patterns, which sharpen as k and ρ increase. In the case of the AR, the fraction of rejections is increasing in the realized value of G_F, regardless of the value of λ_F/k. More weight is progressively assigned to those observations associated to a large first stage F statistic. Instead, the rejection rates for the KLM and MCLR appear U-shaped, as if evidence against the null were correlated with very low (1st decile) and/or very large (10th decile) values of G_F.

Figures 2A and 2B are consistent with the view that the role played by the first stage F statistic in the model studied here is analogous to the role played by the "$X'X$" matrix in the standard linear regression model with random, exogenous regressors. In the linear regression model with random, exogenous regressors (X) whose distribution is known, it is possible to improve on the power of the usual t-test by choosing the critical values as a function of $X'X$. In particular, it is advantageous to have a large conditional size for those designs associated with a small conditional (on $X'X$) variance of the OLS estimator, i.e. designs for which $X'X$ is large (Lehmann (1997) and Le Cam and Yang (2000)). In view of Figures 2A and 2B, a common feature of existing testing procedures is that they achieve unconditional power gains (relative to procedures that do not choose their conditional significance levels as a systematic function of the design) by implicitly choosing conditional size in such a way that the conditional size is large when the first stage F statistic is large.

In light of the figures presented, it is clear that carrying out conditional versus unconditional inference is associated to a significant size-power trade off, reinforcing the importance of identifying the relevant probability calculation for the specific application of interest (see discussion in Section 5).

The second set of numerical results focuses on the two conditional procedures $\phi_\alpha^{CAR}(G;\rho)$ (hereafter conditional AR test or CAR) and the UMPCU test, $\phi_\alpha^{**}(G;\rho)$. In particular, we are interested in evaluating how both tests perform relative to the CPE.

By construction, the power of the CAR test coincides with the CPE as long as $\beta\left(\beta-\sqrt{\omega_{11}/\omega_{22}}\rho\right)$ is positive, but how does the CAR perform relative to the CPE when $\beta\left(\beta-\sqrt{\omega_{11}/\omega_{22}}\rho\right)<0$? Figures 3A, 3B, 3C, and 3D illustrate how the shortcoming of the CAR varies with k, ρ and λ_F/k respectively.

In Figure 3A, the discrepancy between CAR and CPE appears to be increasing in ρ and k, although the difference between the two power curves is barely visible for ρ less than 0.5. Figures 3B and 3C present analogous plots to those in Figure 3A, holding k fixed, and considering a finer grid along the ρ dimension. The discrepancy remains small, even for $\rho = 0.75$. The shortcoming of the CAR is increasing in the normalized measure of the strength of the instruments λ_F/k (see Figure 3D).

For larger values of ρ and λ_F/k, the power of the conditional AR appears to fall below 5%, and therefore CAR is not an unbiased test. The power corresponding to values of β between zero and ρ tends to zero as ρ approaches one. By restricting the class of tests to unbiased testing procedures, it is possible to obtain a UMP (conditionally unbiased) test, as shown in Theorem 4.1. However, the maximum attainable power for values of the alternative between zero and ρ is not large (see CPE for $\beta\in(0,\rho)$), therefore the gains from imposing unbiasedness could be outweighed by losses in power (over different ranges of the alternative) induced by imposing an additional restriction.

Figures 4A and 4B show that there is a nontrivial loss of power when we focus on unbiased testing procedures. Especially for intermediate values of k and ρ, the power associated to values of β in $(0, \rho)$ is virtually flat, and the gains in that region of the parameter space are outweighted by the loss of power everywhere else. We believe that Figures 4A and 4B are representative of the behavior of the power for moderate k. In these cases, the conditional AR exhibit a better overall performance and we recommend it for empirical practice. However, because the attainable power when β ranges between zero and ρ is increasing in the number of instruments, it could be less clear which procedure should be used when k is large.[32]

8 Conclusions

Starting from a generalization of the notion of ancillarity, this paper introduces the class of conditionally similar tests in the context of the endogenous regressor model with homoskedastic Gaussian errors and known covariance matrix. We obtain efficiency bounds (power envelopes) for conditionally similar and conditionally unbiased tests on the coefficient of the endogenous regressor and derive tests that attain these power envelopes.

Making use of weak instruments asymptotics, we show that the efficiency bounds derived under the assumptions Gaussian errors and known covariance matrix can also be attained (asymptotically) when the reduced form covariance matrix has to be estimated and the errors are nonnormal. Furthermore, we show that the limiting power functions obtained under Staiger and Stock (1997) asymptotics are Gaussian asymptotic power envelopes when hypotheses of uncorrelated normal homoskedastic errors and fixed instruments are met, proving that the normality assumption is least favorable.

This paper emphasizes the importance of choosing between conditional and unconditional inference when doing inference on causal relationships. The relevance of the distinction between the conditional and unconditional points of view is illustrated by means of a numerical study. We show that carrying out inference conditionally on the first stage F statistic results in a nontrivial loss of unconditional power, regardless of whether identification is strong or weak. On the other hand, existing testing procedures with good unconditional power properties have bad size properties from the conditional point of view. In some important cases, the IV experiment is of a sequential nature in which the first stage determines the design of the experiment while the second stage determines the outcome of the experiment. We argue that in such cases, the conditional point of view is more appropriate; that is, we argue that in such cases the researcher may want to condition on the design of the experiment (i.e. condition on the first stage F statistic).[33]

When conducting conditional inference, the tests derived in this paper enjoy strong optimality

[32] This conjecture is based on the behavior of the CPE when $k = 50$.

[33] A conceptually different argument that delivers the same conclusion is based on casual empiricism. Many practitioners wish to be able to take value of the first stage F statistic into consideration when doing inference, but are rightly concerned that by doing so a "pretest bias" is introduced. By adopting the conditional point of view promoted herein, empiricists can take the value of the first stage F statistic into account without having to worry about comitting (unintentional) data mining.

properties and are recommended for use in empirical work. The conditional Anderson-Rubin test (i.e., the test that rejects for large values of the AR statistic) is recommended if the number of instruments is moderate. When the number of instruments is large, on the other hand, it is less straightforward to make specific recommendations, but equipped with the numerical results of this paper, empiricists can make an informed choice between the conditional Anderson-Rubin test and the most powerful conditionally unbiased test.

The results derived here rely on a finite sample model that is isomorphic to the limiting experiment under weak instruments asymptotics and fixed number of instruments. In models with many instruments, our argument for conditioning on G_F remains valid, but these models may call for different asymptotic approximations such as many (weak) instruments asymptotics (e.g., Bekker (1994), Chao and Swanson (2003, 2005), Donald and Newey (2001), Hahn (2002), Hansen, Hausman, and Newey (2004), and Stock and Yogo (2003)). In future research, we intend to develop optimality results in these settings. Another interesting generalization of our work that we leave for future research is an investigation of the merits of the conditionality restriction and the derivation of conditional optimality results in the context of GMM.

A Auxiliary Lemmas

This appendix collects a number of lemmas that will be used in the proofs of the main results. Lemma A.1 is used in the proof of Theorem 3.1 and shows that, conditional on G_F, G_C is complete for λ_F.

Lemma A.1 *If $l(\cdot)$ is a function such that*

$$E_{0,\lambda_F}[l(G)|G_F] = 0 \qquad \forall \lambda_F > 0,$$

then

$$E_{0,\lambda_F}[l(G)|G_C, G_F] = 0 \qquad \forall \lambda_F > 0.$$

Proof. Define $L(g_C, g_F) = E_{0,\lambda_F}[l(G)|G_C = g_C,\ G_F = g_F]$, where the expectation is independent of λ_F because G_C is (complete and) sufficient for λ_F under H_0 (Moreira (2003) and AMS). By assumption,

$$\begin{aligned} 0 &= E_{0,\lambda_F}[l(G)|G_F = g_F] \\ &= E_{0,\lambda_F}[L(G_C, g_F)|G_F = g_F] \\ &= E_{0,\lambda_F}[L(G_C, g_F)]. \end{aligned}$$

for every $\lambda_F > 0$ and almost every $g_F > 0$, where the second and third equalities follow by the law of iterated expectations.

Because G_C is complete (and sufficient) for λ_F (Moreira (2003) and AMS), $L(g_C, g_F) = 0$ for almost every $(g_C, g_F) \in \mathbb{R}^2_{++}$. ■

The next result is used in the proofs of Corollaries 3.1 and 4.1.

Lemma A.2 *Let Y and X be random variables and suppose $E|Y|^p < \infty$ for some $p > 1$.*
(a) $E(Y|X) \geq 0$ if and only if $E(Yh(X)) \geq 0$ for every $h \in C_b^+(\mathbb{R})$.
(b) $E(Y|X) = 0$ if and only if $E(Yh(X)) = 0$ for every $h \in C_b^+(\mathbb{R})$.

Proof. In both cases the "only if" part follows by the Law of Iterated Expectations.

The proofs of the "if" parts are by contradiction. First, to prove (a), suppose $E(Y|X) \geq 0$ fails. Then $E[Y \cdot \delta(X)] < 0$, where $\delta(X) = 1\{E(Y|X) < 0\}$. To arrive at a contradiction, it suffices to show that if $E[Y \cdot \delta(X)] < 0$ for some $0 \leq \delta(\cdot) \leq 1$, then $E[Y \cdot h(X)] < 0$ for some $h \in C_b^+(\mathbb{R})$.

Because $E[Y \cdot \delta(X)] < 0$ and $E|Y| < \infty$, it follows from the Dominated Convergence Theorem that $\lim_{K\to\infty} E[Y\delta(X)1\{|X| \leq K\}] = E[Y \cdot \delta(X)] < 0$ and $\lim_{K\to\infty} E[|Y|1\{|X| > K\}] = 0$. Therefore, there exist $\varepsilon > 0$ and $K < \infty$ such that

$$\begin{aligned} E[Y\delta(X)1\{|X| \leq K\}] &< -\varepsilon, \\ E[|Y|1\{|X| > K\}] &< \frac{\varepsilon}{4}. \end{aligned}$$

It follows from the modulus inequality, the triangle inequality, and the Hölder inequality that for any $h \in C_b^+(\mathbb{R})$,

$$\begin{aligned}
& |E[Y\delta(X)] - E[Yh(X)]| \\
\leq\ & E[|Y| \cdot |\delta(X) - h(X)| \cdot 1\{|X| \leq K\}] + 2E[|Y| \cdot 1\{|X| > K\}] \sup_{|x| \geq K} |\delta(x) - h(x)| \\
<\ & E[|Y| \cdot |\delta(X) - h(X)| \cdot 1\{|X| \leq K\}] + \frac{\varepsilon}{2} \sup_{|x| \geq K} |\delta(x) - h(x)| \\
\leq\ & [E(|Y|^p)]^{\frac{1}{p}} \cdot [E(|\delta(X) - h(X)|^q 1\{|X| \leq K\})]^{\frac{1}{q}} + \frac{\varepsilon}{2} \sup_{|x| \geq K} |\delta(x) - h(x)|
\end{aligned}$$

where $p^{-1} + q^{-1} = 1$.

Making use of Dudley (2002, Porblem 7.4.2), $E\left(\left|\delta(X) - \tilde{h}(X)\right|^q 1\{|X| \leq K\}\right)$ can be made smaller than $\varepsilon/2$ by choosing $\tilde{h}$ continuous on $[-K, K]$. In order to complete the proof, we extend the function $\tilde{h}(\cdot)$ to a continuous, $[0,1]$-valued function on $\mathbb{R}$. Define $h(x) = \min\left(\max\left\{\tilde{h}(x), 0\right\}, 1\right)$. The function $h(\cdot)$ is continuous on $[-K, K]$ and satisfies

$$E[|Y| \cdot |\delta(X) - h(X)| \cdot 1\{|X| \leq K\}] \leq E\left[|Y| \cdot \left|\delta(X) - \tilde{h}(X)\right| \cdot 1\{|X| \leq K\}\right]$$

since $|\delta(x) - h(x)| \leq \left|\delta(x) - \tilde{h}(x)\right|$, for any $x \in [-K, K]$. Finally let

$$h(x) = \begin{cases} h(K) & \forall x > K \\ h(-K) & \forall x < -K, \end{cases}$$

which guarantees that $h(\cdot)$ is continuous and $[0,1]$-valued.

Because $0 \leq \delta(\cdot) \leq 1$ and $0 \leq h(\cdot) \leq 1$, $\sup_{|x| \geq K} |\delta(x) - h(x)| \leq 1$. As a consequence,

$$\begin{aligned}
|E[Y\delta(X)] - E[Yh(X)]| &\leq [E(|Y|^p)]^{\frac{1}{p}} \cdot [E(|\delta(X) - h(X)|^q 1\{|X| \leq K\})]^{\frac{1}{q}} + \frac{\varepsilon}{2} \sup_{|x| \geq K} |\delta(x) - h(x)| \\
&< \varepsilon,
\end{aligned}$$

implying

$$E[Yh(X)] \leq E[Y\delta(X)] + |E[Y\delta(X)] - E[Yh(X)]| < 0,$$

completing the proof of part (a).

The proof of the "if" part of (b) is similar. If $E(Y|X) \neq 0$, then $E(Y) \neq 0$ or $E[Y \cdot \delta(X)] < 0$. If $E(Y|X) = 0$ fails because $E(Y) \neq 0$, then $E[Yh(X)] = 0$ fails for $h(\cdot) = 1$. Otherwise, if $E(Y|X) = 0$ fails because $E[Y \cdot \delta(X)] < 0$, then it follows from the proof of part (a) that $E[Y \cdot h(X)] < 0$ for some $h \in C_b^+(\mathbb{R})$. ■

Remark. Lemma A.2 (b) is a special case of Theorem 1 of Bierens (1982). It does not seem possible to base a proof of Lemma A.2 (a) on Bierens (1982, Theorem 1) or its proof.

Lemma A.3 *The conditional null density of G_{AR} given (G_C, G_F) is*

$$f^0_{G_{AR}|G_C,G_F}(g_{AR}|g_C,g_F;\rho) = \frac{\left[g_{AR}g_C - \left(\frac{g_F-\rho^2 g_{AR}-(1-\rho^2)g_C}{2\rho\sqrt{1-\rho^2}}\right)^2\right]^{(k-3)/2}\exp\left\{-\frac{1}{2}(g_{AR})\right\}}{\int_{\underline{g_{AR}}(g_C,g_F;\rho)}^{\overline{g_{AR}}(g_C,g_F;\rho)}\left[tg_C - \left(\frac{g_F-\rho^2 t-(1-\rho^2)g_C}{2\rho\sqrt{1-\rho^2}}\right)^2\right]^{(k-3)/2}\exp\left\{-\frac{1}{2}(t)\right\}dt}, \quad (26)$$

where $\underline{g_{AR}}(g_C,g_F;\rho) = \rho^{-2}\left[\sqrt{(1-\rho^2)g_C} - \sqrt{g_F}\right]^2$ *and* $\overline{g_{AR}}(g_C,g_F;\rho) = \rho^{-2}\left[\sqrt{(1-\rho^2)g_C} + \sqrt{g_F}\right]^2$ *are the lower and upper endpoints of the support, respectively.*

Proof. Define the conditional null density as

$$f^0_{G_{AR}|G_C,G_F}(g_{AR}|g_C,g_F;\rho) = \frac{f_G(g_{AR},g_C,g_F;0,\lambda_F)}{\int f_G(t,g_C,g_F;0,\lambda_F)\,dt},$$

where $\lambda_F > 0$ is arbitrary (and drops out because G_C is sufficient for λ_F under the null). Some algebra delivers the expression in (26). The support of the conditional density of G_{AR} given (G_C, G_F) is determined by the requirement that the matrix in (27) is positive definite, which holds if and only if

$$g_{AR}g_C - \left(\frac{g_F-\rho^2 g_{AR}-(1-\rho^2)g_C}{2\rho\sqrt{1-\rho^2}}\right)^2 > 0.$$

■

Remark. Below, we suppress the dependence of $f^0_{G_{AR}|G_C,G_F}(\cdot)$, $\underline{g_{AR}}(\cdot)$, $\overline{g_{AR}}(\cdot)$ on ρ whenever this does not cause confusion.

Lemma A.3 will be used repeatedly in the sequel. For instance, it is used in the proof of the next result, which itself is used in the proof of Theorem 6.1.

Lemma A.4 *Let $c_{1,\alpha}(\cdot)$ and $c_{2,\alpha}(\cdot)$ be (implicitly) defined such that $\phi^{**}_\alpha(\cdot)$ satisfies (16) and (17). Then $c_{1,\alpha}(\cdot)$ and $c_{2,\alpha}(\cdot)$ are continuous.*

Proof. Continuity is proved by means of the Implicit Function Theorem. Consider the function $F = (F_1, F_2)'$, whose components are

$$F_1(c_1,c_2,\theta) = \int_{c_1}^{c_2} f^0_{G_{AR}|G_C,G_F}(g_{AR}|\theta)\,dg_{AR} - (1-\alpha),$$

$$F_2(c_1,c_2,\theta) = \int_{c_1}^{c_2} g_{AR} f^0_{G_{AR}|G_C,G_F}(g_{AR}|\theta)\,dg_{AR} - (1-\alpha)E_0[G_{AR}|\theta],$$

where $\theta = (g_C, g_F; \rho)$, $c_i \in \left(\underline{g_{AR}}(\theta), \overline{g_{AR}}(\theta)\right)$ (for $i = 1, 2$), and

$$E_0[G_{AR}|\theta] = \int_{\underline{g_{AR}}(\theta)}^{\overline{g_{AR}}(\theta)} g_{AR} f^0_{G_{AR}|G_C,G_F}(g_{AR}|\theta)\, dg_{AR}.$$

For any given value of θ, $c_{1,\alpha}(\theta)$ and $c_{2,\alpha}(\theta)$ are defined such that (16) and (17) hold. Therefore

$$F(c_{1,\alpha}(\theta), c_{2,\alpha}(\theta), \theta) = 0.$$

By Lemma A.3, $\underline{g_{AR}}(\theta)$, $\overline{g_{AR}}(\theta)$, and $f^0_{G_{AR}|G_C,G_F}(g_{AR}, \theta)$ are continuously differentiable in θ. In addition, each component of $F(c_1, c_2, \theta)$ is also continuously differentiable in (c_1, c_2).

Making use of the fact that $c_i \in \left(\underline{g_{AR}}(\theta), \overline{g_{AR}}(\theta)\right)$, it is possible to compute the derivative under the integral sign and it follows from the properties of $f^0_{G_{AR}|G_C,G_F}(\cdot)$ that $\int_{c_1}^{c_2} g_{AR} f^0_{G_{AR}|G_C,G_F}(g_{AR}|\theta)\, dg_{AR}$ is continuously differentiable in θ.

To invoke the Implicit Function Theorem and prove continuity of the critical value function, we need to show that $E_0[G_{AR}|\,\theta]$ is C^1 in θ, i.e. that the derivatives of $E_0[G_{AR}|\theta]$ with respect to each component of θ exist and are continuous.

In light of the definition of the null conditional density, $f^0_{G_{AR}|G_C,G_F}(\cdot)$ (see equation (26)), we consider two distinct cases.

When $k = 2$, $f^0_{G_{AR}|G_C,G_F}(g_{AR}|\theta)$ goes to infinity as g_{AR} approaches $\underline{g_{AR}}(\theta)$ and/or $\overline{g_{AR}}(\theta)$. To show that $E_0[G_{AR}|\cdot] \in C^1$, it is enough to prove that $\int_{\underline{g_{AR}}(\cdot)}^{\overline{g_{AR}}(\cdot)} D(t, \cdot)\, dt \in C^1$, where

$$D(t, \theta) = t\left[t g_C - \left(\frac{\left[g_F - \rho^2 t - (1-\rho^2) g_C\right]}{2\rho\sqrt{1-\rho^2}}\right)^2\right]^{-1/2} \exp\left\{-\frac{1}{2}t\right\}.$$

By Lemma A.3, the change of variables $t = \left(\overline{g_{AR}}(\theta) - \underline{g_{AR}}(\theta)\right)x + \underline{g_{AR}}(\theta)$, and the Dominated Convergence Theorem, $\int_{\underline{g_{AR}}(\cdot)}^{\overline{g_{AR}}(\cdot)} D(t, \cdot)\, dt \in C^1$. (The Dominated Convergence Theorem is applicable because $\int_0^1 [x(1-x)]^{-1/2}\, dx = \pi < \infty$.)

If $k > 2$, then $f^0_{G_{AR}|G_C,G_F}(\cdot|\theta)$ is bounded on $\left(\underline{g_{AR}}(\theta),\ \overline{g_{AR}}(\theta)\right)$, and it follows from the Leibniz Rule, the Dominated Convergence Theorem, and the properties of $f^0_{G_{AR}|G_C,G_F}(\cdot)$, that $E_0[G_{AR}|\cdot] \in C^1$.

The proof is completed by verifying that the Jacobian $J(\theta)$ of partial derivatives of $F(c_1, c_2, \theta)$ with respect to c_1 and c_2 evaluated at $(c_{1,\alpha}(\theta), c_{2,\alpha}(\theta), \theta)$ is non-singular:

$$\begin{aligned} |J(\theta)| &= \left| \left[\begin{array}{cc} \frac{\partial F_1(c_1,c_2,\theta)}{\partial c_1} & \frac{\partial F_1(c_1,c_2,\theta)}{\partial c_2} \\ \frac{\partial F_2(c_1,c_2,\theta)}{\partial c_1} & \frac{\partial F_2(c_1,c_2,\theta)}{\partial c_2} \end{array}\right]\Bigg|_{(c_1,c_2)=(c_{1,\alpha}(\theta),c_{2,\alpha}(\theta))} \right| \\ &= f^0_{G_{AR}|G_C,G_F}(c_{1,\alpha}(\theta), \theta)\, f^0_{G_{AR}|G_C,G_F}(c_{2,\alpha}(\theta), \theta) \cdot [c_{2,\alpha}(\theta) - c_{1,\alpha}(\theta)] \neq 0. \end{aligned}$$

■

The next result will be used in the proof of Theorem 6.2.

Lemma A.5 *Suppose* $\beta \in \mathbb{R}$, $D \in \mathbb{R}^k \backslash \{0\}$, *and* $D_F \in \mathbb{R}^k \backslash \{0\}$. *For any* n, *let* $\mu_n^{0,D}$ *and* μ_n^{β,D_F} *be the probability measures associated with the densities* $f_G\left(\cdot;0,D'\frac{Z'Z}{n}D\right)$ *and* $f_G\left(\cdot;\beta,D_F'\frac{Z'Z}{n}D_F\right)$, *respectively. If* $n^{-1}Z'Z \to Q_{ZZ}$, *a positive definite matrix, then the sequences* $\left\{\mu_n^{0,D}\right\}$ *and* $\left\{\mu_n^{\beta,D_F}\right\}$ *are mutually contiguous.*

Proof. To show that $\left\{\mu_n^{\beta,D_F}\right\}$ is contiguous with respect to $\left\{\mu_n^{0,D}\right\}$, suppose $G_n \sim \mathcal{G}\left(0,D'\frac{Z'Z}{n}D\right)$ for every n. Then, because $n^{-1}Z'Z \to Q_{ZZ}$,

$$\frac{f_G\left(G_n;\beta,D_F'\frac{Z'Z}{n}D_F\right)}{f_G\left(G_n;0,D'\frac{Z'Z}{n}D\right)} = \frac{f_G\left(G_n;\beta,D_F'Q_{ZZ}D_F\right)+o_p(1)}{f_G\left(G_n;0,D'Q_{ZZ}D\right)+o_p(1)}$$

$$\xrightarrow{d} \frac{f_G\left(\mathcal{G}\left(0,D'Q_{ZZ}D\right);\beta,D_F'Q_{ZZ}D_F\right)}{f_G\left(\mathcal{G}\left(0,D'Q_{ZZ}D\right);0,D'Q_{ZZ}D\right)},$$

where the equality uses smoothness of f_G and the convergence result uses the continuous mapping theorem. (Here, "smoothness of f_G" is the property that $\frac{\partial}{\partial\lambda_F}f_G\left(g;0,\lambda_F\right)$ and $\frac{\partial}{\partial\lambda_F}f_G\left(g;\beta,\lambda_F\right)$ are continuous in (g,λ_F). For details, see the remark following the proof.) Now, because the support of f_G not depend on the parameters,

$$E\left[\frac{f_G\left(\mathcal{G}\left(0,D'Q_{ZZ}D\right);\beta,D_F'Q_{ZZ}D_F\right)}{f_G\left(\mathcal{G}\left(0,D'Q_{ZZ}D\right);0,D'Q_{ZZ}D\right)}\right] = \int \frac{f_G\left(g;\beta,D_F'Q_{ZZ}D_F\right)}{f_G\left(g;0,D'Q_{ZZ}D\right)} f_G\left(g;0,D'Q_{ZZ}D\right)dg = 1.$$

By Le Cam's first lemma (van der Vaart (1998, Lemma 6.4) and Hall and Loynes (1977)), $\left\{\mu_n^{\beta,D_F}\right\}$ is contiguous with respect to $\left\{\mu_n^{0,D}\right\}$.

A symmetric argument shows that $\left\{\mu_n^{0,D}\right\}$ is contiguous with respect to $\left\{\mu_n^{\beta,D_F}\right\}$. ■

Remark. For completeness, we now show that

$$\frac{f_G\left(G_n;\beta,D_F'\frac{Z'Z}{n}D_F\right)}{f_G\left(G_n;0,D'\frac{Z'Z}{n}D\right)} = \frac{f_G\left(G_n;\beta,D_F'Q_{ZZ}D_F\right)+o_p(1)}{f_G\left(G_n;0,D'Q_{ZZ}D\right)+o_p(1)}$$

under the conditions of the lemma. Let $\varepsilon > 0$ be given. For any $\delta > 0$ and any set $K \subseteq \mathbb{R}^3$,

$$\begin{aligned}
&\Pr\left[\left|f_G\left(G_n;0,D'\frac{Z'Z}{n}D\right) - f_G\left(G_n;0,D'Q_{ZZ}D\right)\right| > \delta\right] \\
\leq\ & \Pr\left(G_n \notin K\right) + 1\left[\sup_{g\in K}\left|f_G\left(g;0,D'\frac{Z'Z}{n}D\right) - f_G\left(g;0,D'Q_{ZZ}D\right)\right| > \delta\right].
\end{aligned}$$

Because $G_n = O_p(1)$, $\overline{\lim}_{n\to\infty}\Pr\left(G_n \notin K\right) < \varepsilon$ for some compact subset K of the support of f_G. Moreover, because $\partial f_G\left(g;0,\lambda_F\right)/\partial\lambda_F$ is continuous in (g,λ_F) and $n^{-1}Z'Z \to Q_{ZZ}$, it follows from the mean value theorem (and boundedness of continuous functions on compacts) that

$$\lim_{n\to\infty}\sup_{g\in K}\left|f_G\left(g;0,D'\frac{Z'Z}{n}D\right)-f_G\left(g;0,D'Q_{ZZ}D\right)\right|\to 0.$$

As a consequence, $f_G\left(G_n;0,D'\frac{Z'Z}{n}D\right)=f_G\left(G_n;0,D'Q_{ZZ}D\right)+o_p\left(1\right)$.

An analogous argument shows that $f_G\left(G_n;\beta,D'\frac{Z'Z}{n}D\right)=f_G\left(G_n;\beta,D'Q_{ZZ}D\right)+o_p\left(1\right)$.

B Proofs

Proof of Lemma 2.1. Denote the maximal invariant proposed by AMS by

$$Q = \begin{bmatrix} Q_S & Q_{ST} \\ Q_{ST} & Q_T \end{bmatrix}, \tag{27}$$

and let

$$M = \begin{bmatrix} \lambda_{AR} & \delta(\beta)\sqrt{\lambda_{AR}\lambda_C} \\ \delta(\beta)\sqrt{\lambda_{AR}\lambda_C} & \lambda_C \end{bmatrix},$$

where $\lambda_{AR} = \lambda_{AR}(\beta, \lambda_F)$, $\lambda_C = \lambda_C(\beta, \lambda_F)$, and the remaining notation is as in Section 2. By Muirhead (1982, Theorem 10.3.2) and Lemma 3 of AMS, $Q \backsim W_2(k, I_2, M)$; that is, Q is distributed noncentral Wishart with scale matrix I_2, k degrees of freedom, and noncentrality parameter matrix M.

In particular, the density of Q is given by

$$\begin{aligned} f_Q(q;\beta,\lambda) &= K_{AMS}\exp\left(-\tfrac{1}{2}(\lambda_{AR}+\lambda_C)\right)\exp\left(-\tfrac{1}{2}tr(Q)\right)(\det Q)^{(k-3)/2} \\ &\quad \times {}_0\tilde{F}_1\left(;\frac{k}{2};\frac{1}{4}\left[\lambda_{AR}q_S + \lambda_C q_T + 2\delta(\beta)\sqrt{\lambda_{AR}\lambda_C}q_{ST}\right]\right), \end{aligned} \tag{28}$$

where $K_{AMS}^{-1} = 2^k\sqrt{\pi}\Gamma\left(\frac{k-1}{2}\right)$, $\Gamma(\cdot)$ denotes the gamma function, and

$${}_0\tilde{F}_1(;b,z) = \sum_{j=0}^{\infty}\frac{z^j}{\Gamma(b+j)\,j!}$$

is the Regularized Confluent Hypergeometric Function.

(Equation (28) follows from Lemma 3 in AMS and the identity

$${}_0\tilde{F}_1\left(;\frac{k}{2};\frac{1}{4}z\right) = \left(\sqrt{z}\right)^{-\frac{k-2}{2}} I_{\frac{k-2}{2}}\left(\sqrt{z}\right)2^{\frac{k-2}{2}},$$

which relates the modified Bessel Function of the First Kind, $I_{\frac{k-2}{2}}(\cdot)$, to the Regularized (Confluent) Hypergeometric Function, ${}_0\tilde{F}_1(;\cdot;\cdot)$.)

The maximal invariant Q is diffeomorphic to the statistic $G = (G_{AR}, G_C, G_F)$:

$$(Q_S, Q_{ST}, Q_T) = \left(G_{AR}, \frac{1}{2\rho\sqrt{1-\rho^2}}\left[G_F - \rho^2 G_{AR} - \left(1-\rho^2\right)G_C\right], G_C\right).$$

By change of variable and using (28), we obtain the expression stated in (6). ■

Proof of Theorem 3.1. By Lemma A.1 and Lehmann (1997, Theorem 4.2), any conditionally similar test has a Neyman structure with respect to (G_C, G_F). Therefore the relevant density to characterize a point optimal (invariant) conditionally similar test is the conditional density of G_{AR} given (G_C, G_F):

$$f_{G_{AR}|G_C,G_F}\left(g_{AR}|g_{C,}g_F;\beta,\lambda_F\right)=\frac{f_G\left(g_{AR},g_C,g_F;\beta,\lambda_F\right)}{\int_{\mathbb{R}_+}f_G\left(t,g_C,g_F;\beta,\lambda_F\right)dt}.$$

Consider the following testing problem

$$H_0:\beta=0 \quad vs. \quad H_1:\beta=\beta_1 \quad and \quad \lambda_F=\lambda_F^1>0. \tag{29}$$

(By Lemma A.3, the concentration parameter λ_F drops out of the conditional density under the null. For this reason, there is no need to specify λ_F under H_0.) By the Neyman-Pearson Lemma and Lemma A.1, the optimal conditionally similar test for the testing problem in (29) rejects for large values of the of Likelihood Ratio test statistic

$$LR\left(g_{AR}|g_C,g_F;\beta_1,\lambda_F^1\right)=\frac{f_{G_{AR}|G_C,G_F}\left(g_{AR}|g_C,g_F;\beta_1,\lambda_F^1\right)}{f^0_{G_{AR}|G_C,G_F}\left(g_{AR}|g_C,g_F\right)}.$$

It turns out that the Likelihood Ratio test statistic is proportional to ${}_0\tilde{F}_1\left(;\frac{k}{2};Z\left(g_{AR};g_C,g_F,\beta_1,\lambda_F^1\right)\right)$, with a positive proportionality constant that does not depend on g_{AR}:

$$LR\left(g_{AR}|g_C,g_F;\beta_1,\lambda_F^1\right)\propto {}_0\tilde{F}_1\left(;\frac{k}{2};Z\left(g_{AR};g_C,g_F,\beta_1,\lambda_F^1\right)\right), \tag{30}$$

where $Z\left(g_{AR};g_C,g_F,\beta_1,\lambda_F^1\right)$ is an affine function of g_{AR}:

$$\begin{aligned} Z\left(g_{AR};g_C,g_F,\beta_1,\lambda_F^1\right) &= \frac{1}{4}\lambda_F^1\left[a\left(\beta_1\right)g_{AR}+b\left(g_C,g_F;\beta_1\right)\right], \\ a\left(\beta_1\right) &= \frac{1}{1-\rho^2}\beta_1\left(\beta_1-\frac{\omega_{12}}{\omega_{22}}\right). \end{aligned} \tag{31}$$

Because ${}_0\tilde{F}_1\left(;\frac{k}{2};\cdot\right)$ is increasing, the point optimal test rejects when $Z\left(G_{AR};G_C,G_F,\beta_1,\lambda_F^1\right)$ is large. If $\beta_1\left(\beta_1-\sqrt{\omega_{11}/\omega_{22}}\rho\right)$ is positive (negative), $Z\left(\cdot;G_C,G_F,\beta_1,\lambda_F^1\right)$ is strictly increasing (decreasing) and the point optimal test rejects for large (small) values of G_{AR}. If $\beta_1=\sqrt{\omega_{11}/\omega_{22}}\rho$, $Z\left(\cdot;G_C,G_F,\beta_1,\lambda_F^1\right)$ is constant and any conditionally similar test is optimal. ■

Proof of Corollary 3.1. Corollary 3.1 follows from Theorem 3.1 and Lemma A.2. ■

Proof of Lemma 4.1. The conditional power function for a test with test function $\phi(\cdot)$ is given by

$$\begin{aligned} \varphi\left(\beta,\lambda_F,g_F;\phi\right) &= E_{\beta,\lambda_F}\left[\phi\left(G\right)|G_F=g_F\right] \\ &= \int\left[\int\phi\left(g\right)f_{G_{AR}|G_CG_F}\left(g_{AR}|g_C,g_F;\beta,\lambda_F\right)dg_{AR}\right]f_{G_C|G_F}\left(g_C|g_F;\beta,\lambda_F\right)dg_C. \end{aligned}$$

Making use of the fact that the random vector G is a continuously differentiable transformation of $Z'Y$, that G_F is a specific ancillary for β, and invoking properties of the exponential family (Lehmann (1997, Theorem 2.9)), we can conclude that the conditional power function $\varphi(\beta, \lambda_F, g_F; \phi)$ is continuously differentiable in β and that the derivative with respect to β can be computed by differentiating under the integral sign.

The continuity of the power function and the definition of a conditionally unbiased test guarantee that a test is conditionally unbiased only if it is conditionally similar. The necessity of (16) now follows from Lemma A.1.

Differentiability of the power function with respect to β and the form of the testing problem in (9) imply that the power function for a conditionally unbiased test must have a minimum at $\beta = 0$ and that

$$\left.\frac{\partial}{\partial\beta}\varphi(\beta, \lambda_F, g_F; \phi)\right|_{\beta=0} = 0 \qquad \forall \lambda_F > 0.$$

Because any conditionally unbiased test is also conditionally similar,

$$\begin{aligned}&\left.\frac{\partial}{\partial\beta}\varphi(\beta, \lambda_F, g_F; \phi)\right|_{\beta=0}\\ =\;& \int\int \phi(g)\left[\left.\frac{\partial}{\partial\beta} f_{G_{AR}|G_C,G_F}(g_{AR}|g_C, g_F; \beta, \lambda_F)\right|_{\beta=0}\right] dg_{AR} f_{G_C|G_F}(g_C|g_F; 0, \lambda_F)\, dg_C.\end{aligned}$$

To further simplify the above expression, we make use of the following two equalities:

$$\begin{aligned}&\left.\frac{\partial}{\partial\beta} f_{G_{AR}|G_C,G_F}(g_{AR}|g_C, g_F; \beta, \lambda_F)\right|_{\beta=0}\\ &= \frac{f_G'(g_{AR}, g_C, g_F; 0, \lambda_F)}{\int f_G(t, g_C, g_F; 0, \lambda_F)\, dt} - \frac{f_G'(g_{AR}, g_C, g_F; 0, \lambda_F)}{\int f_G(t, g_C, g_F; 0, \lambda_F)\, dt} \cdot \int \frac{f_G'(s, g_C, g_F; 0, \lambda_F)}{\int f_G(t, g_C, g_F; 0, \lambda_F)\, dt} ds\end{aligned}$$

and

$$f_G'(g_{AR}, g_C, g_F; 0, \lambda_F) = \left.\frac{\partial}{\partial\beta} f_G(g_{AR}, g_C, g_F; \beta, \lambda_F)\right|_{\beta=0} = (\gamma_0 + \gamma_{AR} g_{AR}) f_G(g_{AR}, g_C, g_F; \beta, \lambda_F),$$

where

$$\begin{aligned}\gamma_0 &= \gamma_0(g_C, g_F, \lambda_F) = \frac{\lambda_F \omega_{12}}{|\Omega|} + \gamma_C g_C + \gamma_F g_F,\\ \gamma_{AR} &= \gamma_{AR}(g_C, \lambda_F) = \frac{{}_0\tilde{F}_1\left(\frac{k}{2}+1; Z_0(g_C; \lambda_F)\right)}{{}_0\tilde{F}_1\left(\frac{k}{2}; Z_0(g_C; \lambda_F)\right)} \frac{\lambda_F}{4} \frac{\rho^2}{1-\rho^2} > 0,\\ \gamma_C &= \gamma_0(g_C, \lambda_F) = \frac{{}_0\tilde{F}_1\left(\frac{k}{2}+1; Z_0(g_C; \lambda_F)\right)\lambda_F}{{}_0\tilde{F}_1\left(\frac{k}{2}; Z_0(g_C; \lambda_F)\right)}\left(-\frac{\omega_{12}}{|\Omega|} + \frac{1}{4}\right),\\ \gamma_F &= \gamma_0(g_C, g_F, \lambda_F) = -\frac{{}_0\tilde{F}_1\left(\frac{k}{2}+1; Z_0(g_C; \lambda_F)\right)}{{}_0\tilde{F}_1\left(\frac{k}{2}; Z_0(g_C; \lambda_F)\right)} \frac{\lambda_F}{4} \frac{1}{1-\rho^2} < 0,\end{aligned}$$

and

$$Z_0(g_C;\lambda_F) = Z(g_{AR};g_C,g_F,0,\lambda_F) = \sqrt{\frac{\omega_{11}}{|\Omega|}}\lambda_F g_C.$$

Therefore,

$$\begin{aligned}
0 &= \left.\frac{\partial}{\partial\beta}\varphi(\beta,\lambda_F,g_F;\phi)\right|_{\beta=0} \\
&= \int\left[\int \phi(g)\, g_{AR} f^0_{G_{AR}|G_C,G_F}(g_{AR}|g_C,g_F)\, dg_{AR} - \right. \\
&\quad \left.\int \phi(g)\, f^0_{G_{AR}|G_C,G_F}(g_{AR}|g_C,g_F)\, dg_{AR} \int g_{AR} f^0_{G_{AR}|G_C,G_F}(g_{AR}|g_C,g_F)\, dg_{AR}\right] \\
&\quad \times f_{G_C|G_F}(g_C|g_F;0,\lambda_F)\, \gamma_{AR} dg_C \\
&= \int \gamma_{AR} Cov_0(\phi(G), G_{AR}|G_C = g_C, G_F = g_F)\ \ f_{G_C|G_F}(g_C|G_F = g_F;0,\lambda_F)\, dg_C,
\end{aligned}$$

where $Cov_0(\phi(G), G_{AR}|G_C, G_F)$ denotes the null covariance between $\phi(G)$ and G_{AR}, conditional on the pair (G_C, G_F).

Since $\gamma_{AR} > 0$ and since G_C given G_F is complete for λ_F (Lemma A.1), $\partial\varphi(\beta,\lambda_F,g_F;\phi)/\partial\beta|_{\beta=0}$ is equal to zero if and only if $Cov_0(\phi(G)\, G_{AR}|G_C = g_C, G_F = g_F) = 0$, completing the proof of the necessity of (17). ■

Proof of Theorem 4.1. The method of proof follows Lehmann (1997, page 136). Consider the following testing problem

$$H_0 : \beta = 0 \quad vs. \quad H_1 : \beta = \beta_1 \quad and \quad \lambda_F = \lambda_F^1 > 0.$$

Let $M(G_C, G_F)$ be the set of points

$$\left(E_{0,\lambda_F}[\phi(G;\rho)|G_C, G_F],\ E_{0,\lambda_F}(\phi(G;\rho)\, G_{AR}|G_C, G_F)\right)$$

as $\phi(\cdot)$ ranges over the class of all test functions. Because the conditional expectation is a linear operator, $M(G_C, G_F)$ is convex and contains all points $(u, uE_{0,\lambda_F}(G_{AR}|G_C, G_F))$ with $u \in (0,1)$. In order to invoke Lehmann (1997, Theorem 3.6(iv)), we need to show that the point $(\alpha, \alpha E_{0,\lambda_F}(G_{AR}|G_C, G_F))$ is in the interior of $M(G_C, G_F)$. That is, we need to show that $M(G_C, G_F)$ also contains the points (α, u_1), with $u_1 > \alpha E_{0,\lambda_F}(G_{AR}|G_C, G_F)$, and (α, u_2), with $u_2 < \alpha E_{0,\lambda_F}(G_{AR}|G_C, G_F)$.

If $(\alpha, u_1) \in M(G_C, G_F)$, then there exists a conditionally similar test that has positive derivative at zero. Consider $\tilde{\phi}_\alpha(G) = 1[G_{AR} > \tilde{c}_\alpha(G_C, G_F)]$, with $\tilde{c}_\alpha(\cdot)$ defined as in (12). Then

$$\begin{aligned}
&\frac{\partial}{\partial\beta}\int \tilde{\phi}_\alpha(g)\, f_{G_{AR}|G_C,G_F}(g_{AR}|G_C, G_F; \beta, \lambda_F)\, dg_{AR}\Big|_{\beta=0} \\
=\;& \gamma_{AR} Cov_0\left(\tilde{\phi}_\alpha(G), G_{AR}|G_C, G_F\right) \\
=\;& \gamma_{AR}\left[\int_{\underline{g_{AR}}(G_C,G_F)}^{\overline{g_{AR}}(G_C,G_F)} \tilde{\phi}_\alpha(g)\, g_{AR} f^0_{G_{AR}|G_C,G_F}(g_{AR}|G_C, G_F)\, dg_{AR} - \alpha E_{0,\lambda_F}(G_{AR}|G_C, G_F)\right] \\
=\;& \alpha\cdot\gamma_{AR}\left[\int_{\tilde{c}_\alpha(G_C,G_F)}^{\overline{g_{AR}}(G_C,G_F)} g_{AR}\frac{f^0_{G_{AR}|G_C,G_F}(g_{AR}|G_C,G_F)}{\Pr_0[G_{AR}>\tilde{c}_\alpha(G_C,G_F)|G_C,G_F]} dg_{AR} - E_{0,\lambda_F}(G_{AR}|G_C, G_F)\right] \\
=\;& \alpha\cdot\gamma_{AR}\left[E_0(G_{AR}|G_C, G_F, G_{AR} > \tilde{c}_\alpha(G_C, G_F)) - E_{0,\lambda_F}(G_{AR}|G_C, G_F)\right] > 0, \qquad (32)
\end{aligned}$$

where the third equality uses $\Pr_0[G_{AR} > \tilde{c}_\alpha(G_C, G_F)|G_C, G_F] = E_{0,\lambda_F}\left[\tilde{\phi}_\alpha(G)|G_C, G_F\right] = \alpha$ and the inequality uses $\gamma_{AR} > 0$ and $\tilde{c}_\alpha(G_C, G_F) > \underline{g_{AR}}(G_C, G_F) \geq 0$.

Similarly, if $M(G_C, G_F)$ contains points (α, u_2), there exists a conditionally similar test that has negative derivative at zero. In fact, consider the test $\hat{\phi}_\alpha(G) = 1[G_{AR} < \hat{c}_\alpha(G_C, G_F)]$, where $\hat{c}_\alpha(\cdot)$ is defined as in (12). Then

$$\begin{aligned}
&\frac{\partial}{\partial\beta}\int \hat{\phi}_\alpha(g)\, f_{G_{AR}|G_C,G_F}(g_{AR}|G_C, G_F; \beta, \lambda_F)\, dg_{AR}\Big| \\
=\;& \gamma_{AR} Cov_0\left(\hat{\phi}_\alpha(G), G_{AR}|G_C, G_F\right) \\
=\;& \gamma_{AR}\int_{\underline{g_{AR}}(G_C,G_F)}^{\hat{c}_\alpha(G_C,G_F)} g_{AR} f^0_{G_{AR}|G_C,G_F}(g_{AR}|G_C, G_F)\, dg_{AR} - \alpha E_{0,\lambda_F}(G_{AR}|G_C, G_F) \\
=\;& \alpha\cdot\gamma_{AR}\left[E_{0,\lambda_F}(G_{AR}|G_C, G_F, G_{AR} < \hat{c}_\alpha(G_C, G_F)) - E_{0,\lambda_F}(G_{AR}|G_C, G_F)\right] < 0. \qquad (33)
\end{aligned}$$

By Lehmann (1997, Theorem 3.5 (iv)), there exist constants k_1, k_2 and a test $\phi^{**}_\alpha(G;\rho)$ satisfying (16) and (17), whose rejection region is

$$LR\left(G_{AR}|G_C, G_F; \beta_1, \lambda^1_F\right) > k_1 + k_2 G_{AR}, \qquad (34)$$

or equivalently

$${}_0\tilde{F}_1\left[\frac{k}{2}; \frac{\lambda_F}{4} a(\beta_1) G_{AR} + b(G_C, G_F; \beta_1)\right] > \tilde{k}_1 + \tilde{k}_2 G_{AR}, \qquad (35)$$

for some $\tilde{k}_1$ and $\tilde{k}_2$ which are functions of β_1, λ^1_F, G_C, and G_F.

The rejection region described by (35) is a subset of $\mathbb{R}$. Because ${}_0\tilde{F}_1\left(\frac{k}{2};\cdot\right)$ is a (strictly) convex function, the subset can be either the entire real line, the empty set, a one sided interval, or a two sided interval.

The set of points for which the left and right hand sides of (35) are equal cannot be an empty set, nor the entire real line. These two scenarios are ruled out by the fact that $\phi_{\alpha}^{**}(G;\rho)$ has to have conditional size α. Also, (35) cannot represent a one sided interval, since it would violate the local to zero derivative restriction in (17). In fact, we have shown that tests which reject for large values of G_{AR} or for small values of G_{AR} do not satisfy (17) (see (33) and (32)).

Therefore, the rejection region in (35) has to be of the form

$$\phi_{\alpha}^{**}(G;\rho) = 1\left[G_{AR} < c_{1,\alpha}(G_C, G_F;\rho)\right] + 1\left[G_{AR} > c_{2,\alpha}(G_C, G_F;\rho)\right], \tag{36}$$

where $c_{1,\alpha}(G_C, G_F;\rho)$ and $c_{2,\alpha}(G_C, G_F;\rho)$ are defined such that (16) and (17) are satisfied.

Finally, because (16) and (17) do not involve β_1 (or λ_F^1), $c_{1,\alpha}(G_C, G_F;\rho)$ and $c_{2,\alpha}(G_C, G_F;\rho)$ the functional form of the critical region in (36) does not depend on the value of the specific alternative $\left(\beta_1, \lambda_F^1\right)$. Therefore, $\phi_{\alpha}^{**}(G;\rho)$ is the UMP conditionally unbiased test. ■

Proof of Corollary 4.1. Corollary 4.1 follows from Theorem 4.1 and Lemma A.2. ■

Proof of Theorem 6.1. Making use of equations (1)-(2) and assumptions A1-A4,

$$\frac{1}{\sqrt{n}}Z'Y = \frac{Z'Z}{n}\sqrt{n}\Pi(\beta,1) + \frac{1}{\sqrt{n}}Z'V \xrightarrow{d} N\left(Q_{ZZ}D(\beta,1), \Omega\otimes Q_{ZZ}\right) \tag{37}$$

and

$$\hat{\Omega} = (Y - N_Z Y)'(Y - N_Z Y)/n \xrightarrow{p} \Omega. \tag{38}$$

Because $\hat{G}$ is a continuous function of $\left(n^{-1/2}Z'Y, \hat{\Omega}\right)$, equations (37)-(38) and the Continuous Mapping Theorem imply

$$\left(\hat{G}_{AR}, \hat{G}_C, \hat{G}_F\right) \xrightarrow{d} \left(\mathcal{G}_{AR}(\beta, \lambda_{F,\infty}), \mathcal{G}_C(\beta, \lambda_{F,\infty}), \mathcal{G}_F(\lambda_{F,\infty})\right). \tag{39}$$

It follows from the definition of $\mathcal{G}(\beta, \lambda_{F,\infty})$ that all the distributional results derived in Lemmas 2.1, A.3 and A.1 apply to the random vector $\left(\mathcal{G}_{AR}(\beta, \lambda_{F,\infty}), \mathcal{G}_C(\beta, \lambda_{F,\infty}), \mathcal{G}_F(\lambda_{F,\infty})\right)$. Therefore, the conditional density of $\mathcal{G}_{AR}(\beta, \lambda_{F,\infty})$ given $\left(\mathcal{G}_C(\beta, \lambda_{F,\infty}), \mathcal{G}_F(\lambda_{F,\infty})\right)$ is absolutely continuous and is a continuous function of $\left(\mathcal{G}_C(\beta, \lambda_{F,\infty}), \mathcal{G}_F(\lambda_{F,\infty}), \rho\right)$. It follows from this property that the critical value function $c_{\alpha}(g_C, g_F;\rho,\beta_1)$, defined in (12), is a continuous function in (g_C, g_F, ρ) (Jansson and Moreira (2004, Lemma 12 (a))). Continuity of the critical value functions $c_{1,\alpha}(\cdot)$ and $c_{2,\alpha}(\cdot)$ defined in Theorem 4.1 can be established invoking Lemma A.4. By the Continuous Mapping Theorem,

$$c_{i,\alpha}\left(\hat{G}_C, \hat{G}_F;\hat{\rho}\right) \xrightarrow{d} c_{i,\alpha}\left(\mathcal{G}_C(0, \lambda_{F,\infty}), \mathcal{G}_F(\lambda_{F,\infty});\rho\right) \qquad i = 0,1,2. \tag{40}$$

where $c_{0,\alpha}(\cdot,\cdot;\cdot)$ is used as short hand for the critical value function $c_{\alpha}(\cdot,\cdot;\cdot,\beta)$ and the convergence holds jointly with (39). Finally, equations (22)-(23) follow by combining (39) and (40) and applying the Continuous Mapping Theorem and Billingsley (1999, Theorem 3.5) ■

Proof of Theorem 6.2. Let β, D, and $\lambda_{F,\infty}$ be given. Let $\Delta_1(\alpha,\rho)$ be the class of test functions $\phi(\cdot)$ that satisfy

$$E\left[\left(\phi\left(\mathcal{G}\left(0,\lambda_F\right)\right)-\alpha\right)h\left(\mathcal{G}_F\left(\lambda_F\right)\right)\right]=0 \qquad \forall\lambda_F>0$$

for every $h\in C_b^{+}\left(\mathbb{R}_{++}\right)$.

It follows from Theorem 3.1 that

$$E\left\{\left[\phi\left(\mathcal{G}\left(\beta,\lambda_{F,\infty}\right)\right)-\phi_{\alpha}^{*}\left(\mathcal{G}\left(\beta,\lambda_{F,\infty}\right);\rho,\beta\right)\right]h\left(\mathcal{G}_F\left(\lambda_{F,\infty}\right)\right)\right\}\leq 0$$

for every $\phi\in\Delta_1(\alpha,\rho)$ and every $h\in C_b^{+}\left(\mathbb{R}_{++}\right)$.

To complete the proof of (a), we show that for any sequence of test functions $\{\phi_n(\cdot)\}$ satisfying (24), there exists a test $\phi\in\Delta_1(\alpha,\rho)$ such that

$$E\left[\phi\left(\mathcal{G}\left(\beta,\lambda_{F,\infty}\right)\right)h\left(\mathcal{G}_F\left(\lambda_{F,\infty}\right)\right)\right]=\overline{\lim}_{n\to\infty}E_{\beta,D}\left[\phi_n(G)\,h\left(G_F\right)\right].$$

Consider an arbitrary sequence of test functions $\{\phi_n(\cdot)\}$ that satisfies (24). Because $\{E_{\beta,D}\left[\phi_n(G)\,h\left(G_F\right)\right]\}$ is bounded, there exists a convergent subsequence such that

$$\lim_{j\to\infty}E_{\beta,D}\left[\phi_{n(j)}(G)\,h\left(G_F\right)\right]=\overline{\lim}_{n\to\infty}E_{\beta,D}\left[\phi_n(G)\,h\left(G_F\right)\right].$$

In addition, since $\phi_{n(j)}$ is bounded in probability, it follows from Prohorov's theorem (van der Vaart (1998, Theorem 2.4)), that there exists a further subsequence such that

$$\left(\phi_{n(j(l))}(G),G\right)\overset{d_0}{\to}\left(\phi_{\infty},\mathcal{G}\left(0,\lambda_{F,\infty}\right)\right),$$

where $\overset{d_0}{\to}$ denotes the convergence in distribution when $(\beta,\lambda_F)=(0,\lambda_{F,\infty})$ and ϕ_{∞} represents some random variable defined on the same probability space as $\mathcal{G}(0, \lambda_{F,\infty})$. Therefore,

$$\begin{aligned}
\overline{\lim}_{n\to\infty} E_{\beta,D}\left[\phi_n(G)\, h(G_F)\right] &= \lim_{j\to\infty} E_{\beta,D}\left[\phi_{n(j)}(G)\, h(G_F)\right] \\
&= \lim_{l\to\infty} E_{\beta,D}\left[\phi_{n(j(l))}(G)\, h(G_F)\right] \\
&= E\left[\phi_\infty h(\mathcal{G}_F(\lambda_{F,\infty})) \cdot \frac{f_G(\mathcal{G}(0,\lambda_{F,\infty});\beta,\lambda_{F,\infty})}{f_G(\mathcal{G}(0,\lambda_{F,\infty});0,\lambda_{F,\infty})}\right] \\
&= E\left[\phi(\mathcal{G}(0,\lambda_{F,\infty}))\, h(\mathcal{G}_F(\lambda_{F,\infty})) \cdot \frac{f_G(\mathcal{G}(0,\lambda_{F,\infty});\beta,\lambda_{F,\infty})}{f_G(\mathcal{G}(0,\lambda_{F,\infty});0,\lambda_{F,\infty})}\right] \\
&= \int \phi(g)\, h(g_F) \cdot \frac{f_G(g;\beta,\lambda_{F,\infty})}{f_G(g;0,\lambda_{F,\infty})} f_G(g;0,\lambda_{F,\infty})\, dg \\
&= \int \phi(g)\, h(g_F) \cdot f_G(g;\beta,\lambda_{F,\infty})\, dg \\
&= E\left[\phi(\mathcal{G}(\beta,\lambda_{F,\infty});\rho,\beta)\, h(\mathcal{G}_F(\lambda_{F,\infty}))\right],
\end{aligned}$$

where $\phi(\mathcal{G}(0,\lambda_{F,\infty})) = E[\phi_\infty | \mathcal{G}(0,\lambda_{F,\infty})]$, the third equality follows from Lemma A.5 (and its proof), Le Cam's Third Lemma and Billingsley (1999, Theorem 3.5), and the fourth equality uses the Law of Iterated Expectations.

The proof is completed by showing that $\phi \in \Delta_1(\alpha,\rho)$. For any $\lambda_F > 0$,

$$\begin{aligned}
& E\left([\phi(\mathcal{G}(0,\lambda_F)) - \alpha]\, h(\mathcal{G}_F(\lambda_F))\right) \\
= {} & E\left([\phi(\mathcal{G}(0,\lambda_{F,\infty})) - \alpha]\, h(\mathcal{G}_F(\lambda_{F,\infty})) \cdot \frac{f_G(\mathcal{G}(0,\lambda_{F,\infty});0,\lambda_F)}{f_G(\mathcal{G}(0,\lambda_{F,\infty});0,\lambda_{F,\infty})}\right) \\
= {} & E\left([\phi_\infty - \alpha]\, h(\mathcal{G}_F(\lambda_{F,\infty})) \cdot \frac{f_G(\mathcal{G}(0,\lambda_{F,\infty});\beta,\lambda_F)}{f_G(\mathcal{G}(0,\lambda_{F,\infty});0,\lambda_{F,\infty})}\right) \\
= {} & \lim_{l\to\infty} E_{0,D_F}\left\{\left[\phi_{n(j(l))}(G) - \alpha\right] h(G_F)\right\} = 0,
\end{aligned}$$

where $D_F \in \mathbb{R}^k$ is such that $D_F' Q_{ZZ} D_F = \lambda_F$, the first equality uses $\phi(\mathcal{G}(0,\lambda_{F,\infty})) = E[\phi_\infty|\mathcal{G}(0,\lambda_{F,\infty})]$, the second equality follows from Lemma A.5, Le Cam's Third Lemma and Billingsley (1999, Theorem 3.5), while the third equality uses the assumption that $\{\phi_n(\cdot)\}$ satisfies (24).

The method of proof of part (b) is similar to that of part (a). It follows from Theorem 4.1 that

$$E\left\{\left[\phi(\mathcal{G}(\beta,\lambda_{F,\infty})) - \phi_\alpha^{**}(\mathcal{G}(\beta,\lambda_{F,\infty});\rho)\right] h(\mathcal{G}_F(\lambda_{F,\infty}))\right\} \leq 0 \qquad \forall \phi(\cdot) \in \Delta_2(\alpha,\rho),$$

where $\Delta_2(\alpha,\rho)$ is the class of test functions $\phi(\cdot)$ that satisfy

$$\begin{aligned}
E\left[(\phi(\mathcal{G}(0,\lambda_F)) - \alpha)\, h(\mathcal{G}_F(0,\lambda_F))\right] &= 0 \qquad \forall \lambda_F > 0 \\
E\left[(\phi_\alpha^{**}(\mathcal{G}(0,\lambda_F);\rho) - \alpha)\, \mathcal{G}_{AR}(0,\lambda_F)\, h(\mathcal{G}_F(0,\lambda_F))\right] &= 0 \qquad \forall \lambda_F > 0
\end{aligned}$$

for every $h \in C_b^+(\mathbb{R}_{++})$. Proceeding as in the proof of part (a), the proof of part (b) is completed by showing that there is a test function $\phi(\cdot)$ in $\Delta_2(\alpha,\rho)$ such that

$$\overline{\lim}_{n\to\infty} E_{\beta,D}\left[\phi_n(G)\, h(G_F)\right] = E\left[\phi\left(\mathcal{G}\left(\beta,\lambda_{F,\infty}\right)\right) h\left(\mathcal{G}_F\left(\lambda_{F,\infty}\right)\right)\right].$$

■

C Appendix - Conditioning

In this section, we make the parallel between the Cox-type example discussed in Section 5 and the IV model explicit. In particular, we show that corresponding to the IV problem there is a Cox-type example in which two tosses of a biased coin (with unknown probability of observing heads or tails) are performed to determine which experiment to run, but only one toss is observed. In the IV model, the object that corresponds to the observed coin toss is the first stage F statistic. We therefore argue that, in some cases, conditioning on the observed value of the first stage F statistic helps to select that portion of the sample space that is relevant to the problem at hand (in the same way that conditioning on the observed coin toss does in the aforementioned Cox-type example).

C.1 Embedding Model, Partial Ancillarity, and Experimental Mixtures

Our discussion is organized as follows. We first show that the model in Section 2.1 can be embedded in a latent model that admits a partial ancillary (Basu (1977)). Then we argue that conditioning on a partial ancillary is appropriate if the sampling scheme underlying the statistical model can be thought of as an experimental mixture (Kalbfleisch (1975)). Finally, we exploit the fact that the partial ancillary in the latent model nests the specific ancillary in the original model (G_F) and conclude that we should condition on G_F since it is the only part of the partial ancillary that is available to us.

The debate on whether or not to condition is largely centered around the notion of partial ancillarity and most of the cases in which conditioning is considered legitimate involve models which admit partial ancillaries.[34] In order to relate our results to the existing literature, we therefore show that the model in Section 2.1 can be embedded in a latent model that admits a partial ancillary, which nests the specific ancillary in the original model (G_F).

Consider the model in Section 2.1 and assume that the endogeneity arises because of an omitted variable, say y_3. Specifically, consider the latent model represented by the following three equations:

$$\begin{aligned} y_1 &= y_2\beta + y_3\gamma + \varepsilon, \\ y_2 &= Z\Pi + v_2, \\ y_3 &= v_3, \end{aligned}$$

where y_1, y_2 and y_3 are n-dimensional vectors, Z is an $n \times k$ full column rank matrix of nonrandom regressors, and each row of the $n \times 3$ matrix of disturbances $[\varepsilon : v_2 : v_3]$ is assumed to be iid (trivariate) normal with mean zero and a covariance matrix $\bar{\Sigma}$, which is block diagonal when partitioned after the first row and column:[35]

$$\bar{\Sigma} = \begin{bmatrix} \sigma_{\varepsilon\varepsilon} & 0 & 0 \\ 0 & \omega_{22} & \omega_{23} \\ 0 & \omega_{23} & \omega_{33} \end{bmatrix}.$$

[34] Among others, see Cox (1988), Kalbfleisch (1975), Lehmann (1997) and Lehmann and Scholz (1992).

[35] Under the assumptions of Section 2, it is always possible to augment the original model in this way.

The model can be written in reduced form as

$$\begin{aligned} y_1 &= Z\Pi\beta + v_1, \\ y_2 &= Z\Pi + v_2, \\ y_3 &= v_3, \end{aligned}$$

where $v_1 = \varepsilon + V\theta$, $\theta = (\beta, \gamma)'$ and $V = [v_2 : v_3]$. The covariance matrix for the i-th row of $[v_1 : v_2 : v_3]$ is denoted by

$$\bar{\Omega} = E\left[(v_{i1}, V_i)'(v_{i1}, V_i)\right] = \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} \\ \omega_{21} & \omega_{22} & \omega_{23} \\ \omega_{31} & \omega_{32} & \omega_{33} \end{bmatrix}.$$

As before, β is the object of interest and the k-dimensional vector Π is an unknown nuisance parameter. The parameters ω_{23}, ω_{33}, and $\Omega = E\left[(v_{i1}, v_{i2})'(v_{i1}, v_{i2})\right]$ are assumed to be known.[36,37] The parameters β and Π are variation independent (i.e., β and Π are free to vary in their domains).

The property that ε_i is independent of $(v_{2i}, v_{3i})'$, is a requirement on y_{3i}. It captures the control variable idea (see Blundell and Powell (2004)) that u_i, error term of the structural equation in (1), depends on $(y_{2i}, y_{3i})'$ only through $y_{3i} = v_{3i}$. As in the Cox example, the variation independence assumption implies that knowing Π, the parameter that governs the (marginal) distribution of $\left\{(y_{2i}, y_{3i})'\right\}$, does not restrict the range of β, the parameter that governs the conditional distribution of $\{y_{1i}\}$ given $\left\{(y_{2i}, y_{3i})'\right\}$ (and vice versa). By implication, $\left\{(y_{2i}, y_{3i})'\right\}$ is a partial ancillary for β.

In Sections 2.1 and 2.2, sufficiency and invariance arguments enabled us to reduce the data to the tridimensional vector (G_{AR}, G_C, G_F), where G_F is a specific ancillary for β. In the latent model, the counterpart of (G_{AR}, G_C, G_F) is given by the six-dimensional vector

$$\left(\tilde{y}_1'\tilde{y}_1', \tilde{y}_1'\tilde{y}_2, \tilde{y}_1'\tilde{y}_3, vech\left(\tilde{Y}'\tilde{Y}\right)\right),$$

where $\tilde{y}_i = (Z'Z)^{-1/2} Z'y_i$ (for $i = 1, 2, 3$) and $\tilde{Y} = [\tilde{y}_2 : \tilde{y}_3]$.

It can be shown that $vech\left(\tilde{Y}'\tilde{Y}\right)$ is a partial ancillary for β. As a result, the likelihood for $\left(\tilde{y}_1'\tilde{y}_1, \tilde{y}_1'\tilde{y}_2, \tilde{y}_1'\tilde{y}_3, vech\left(\tilde{Y}'\tilde{Y}\right)\right)$ can be factorized as follows:

$$f\left(\tilde{y}_1'\tilde{y}_1, \tilde{y}_1'\tilde{y}_2, \tilde{y}_1'\tilde{y}_3, vech\left(\tilde{Y}'\tilde{Y}\right); \beta, \lambda_F\right) \propto f\left(\tilde{y}_1'\tilde{y}_1, \tilde{y}_1'\tilde{y}_2, \tilde{y}_1'\tilde{y}_3 | vech\left(\tilde{Y}'\tilde{Y}\right); \beta\right) f\left(vech\left(\tilde{Y}'\tilde{Y}\right); \lambda_F\right),$$

[36] Under the stated assumptions, γ and ω_{13} can be expressed as functions of the parameters in the model:

$$\begin{aligned} \gamma &= \gamma(\beta; \omega_{32}, \Omega) = \omega_{23}^{-1}(\omega_{12} - \beta\omega_{22}), \\ \omega_{13} &= \omega_{13}(\beta; \omega_{32}, \omega_{33}, \Omega) = \omega_{23}\beta + \omega_{33}\gamma(\beta; \omega_{32}, \Omega). \end{aligned}$$

It is important to notice that γ does not depend on Π. Moreover, the fact ω_{13} does not depend on Π implies that $\sigma_{\varepsilon\varepsilon} = Var(v_{1i}|v_{2i}, v_{3i})$ does not depend on Π.

[37] An argument analogous to that made here goes through if the reduced form covariance matrix Ω is treated as an unknown nuisance parameter and/or normality assumption is dropped. For specificity, we consider only the case correponding to the model studied in the main text. Details about more general cases are available from the authors upon request.

where $f(\cdot)$ denotes the relevant density identified by its argument. Moreover, because G_F is proportional to $\tilde{y}_2'\tilde{y}_2$, the partial ancillary $vech\left(\tilde{Y}'\tilde{Y}\right)$ nests the specific ancillary (G_F) in the original model.

Although conditioning on partial ancillaries is very tempting (and common practice), the choice to condition cannot be justified solely on the basis of a convenient factorization of the likelihood. Rather, it is important to identify whether the repeated sampling underlying the model of interest corresponds to draws from the conditional or the joint distribution (Cox (1988), Kalbfleisch (1975), Lehmann (1997) and Lehmann and Scholz (1992)). As a consequence, it is useful to characterize the statistical models in terms of sampling schemes.

Given the factorization of the likelihood associated to the definition of partial ancillary, it is always possible to think of the data as being generated according to a mixture model (Birnbaum (1962) and Lehmann (1997)). Kalbfleisch (1975) and Basu (1964) have distinguished between experimental and mathematical mixtures. The former results from the "experimental design" under study, whereas the latter is a product of the mathematical properties of the hypothesized parametric model. This distinction is particularly informative in the case of partial ancillaries. In fact, in order to justify the decision to condition, the factorization of the likelihood has to be a reflection of the actual experimental design (or statistical model under study), not a simple mathematical virtue of the statistical model. In the endogenous regressor model it is therefore important that the embedding model is not simply a mathematical artifact, but rather corresponds (as does the model above) to a latent model which explains the source of endogeneity in the original model.

In the omitted variable case, the data can be thought of as generated in three stages. First, the experimental design ($vech\left(\tilde{Y}'\tilde{Y}\right)$) is selected. Then, conditional on $vech\left(\tilde{Y}'\tilde{Y}\right)$, the outcomes $(\tilde{y}_1'\tilde{y}_1, \tilde{y}_1'\tilde{y}_2, \tilde{y}_1'\tilde{y}_3)$ are generated. Finally, part of the information concerning the experiment (the parts involving y_3, namely $\tilde{y}_1'\tilde{y}_3$, $\tilde{y}_2'\tilde{y}_3$, and $\tilde{y}_3'\tilde{y}_3$) is lost in the process of collecting the data. Even though we are able to observe only a subset of the information contained in the partial ancillary $vech\left(\tilde{Y}'\tilde{Y}\right)$, the repeated sampling scheme underlying the model of interest does not change and it seems desirable to condition on the available information concerning the experimental design, namely $\tilde{y}_2'\tilde{y}_2$ (which differs from G_F only by an unimportant scale factor). In this sense, the data underlying the IV model can be thought of as being generated by a Cox-type example in which two tosses of a coin (with unknown probabilities of observing heads or tails) determine which experiment to run, but only one toss is observed. Specifically, the observed coin toss corresponds to $\tilde{y}_2'\tilde{y}_2$, while the experiment generates the observed values of $\tilde{y}_1'\tilde{y}_1$ and $\tilde{y}_1'\tilde{y}_2$. As in the original Cox example, the relevant probability calculations would appear to be those based on the conditional distribution of $\tilde{y}_1'\tilde{y}_1$, $\tilde{y}_1'\tilde{y}_2$, given $\tilde{y}_2'\tilde{y}_2$. Equivalently, the relevant probability calculations would appear to be those based on the conditional distribution of (G_{AR}, G_C) given G_F. (The equivalence between the two statements follows from simple algebra.)

D - Appendix - Figures

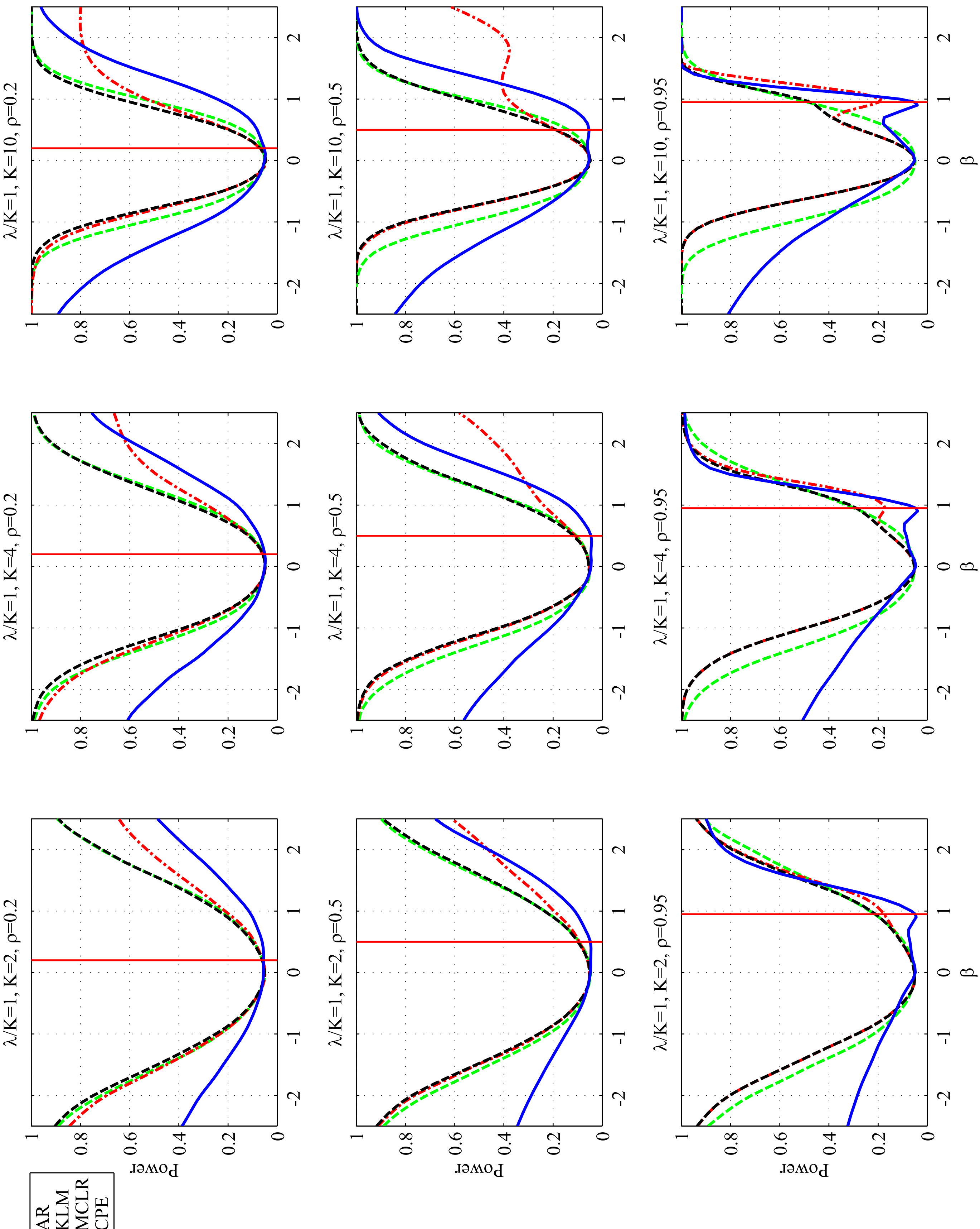

Figure 1A: Unconditional Power Plots. Different values of k and ρ (vertical red line corresponds to $\beta = \rho$).

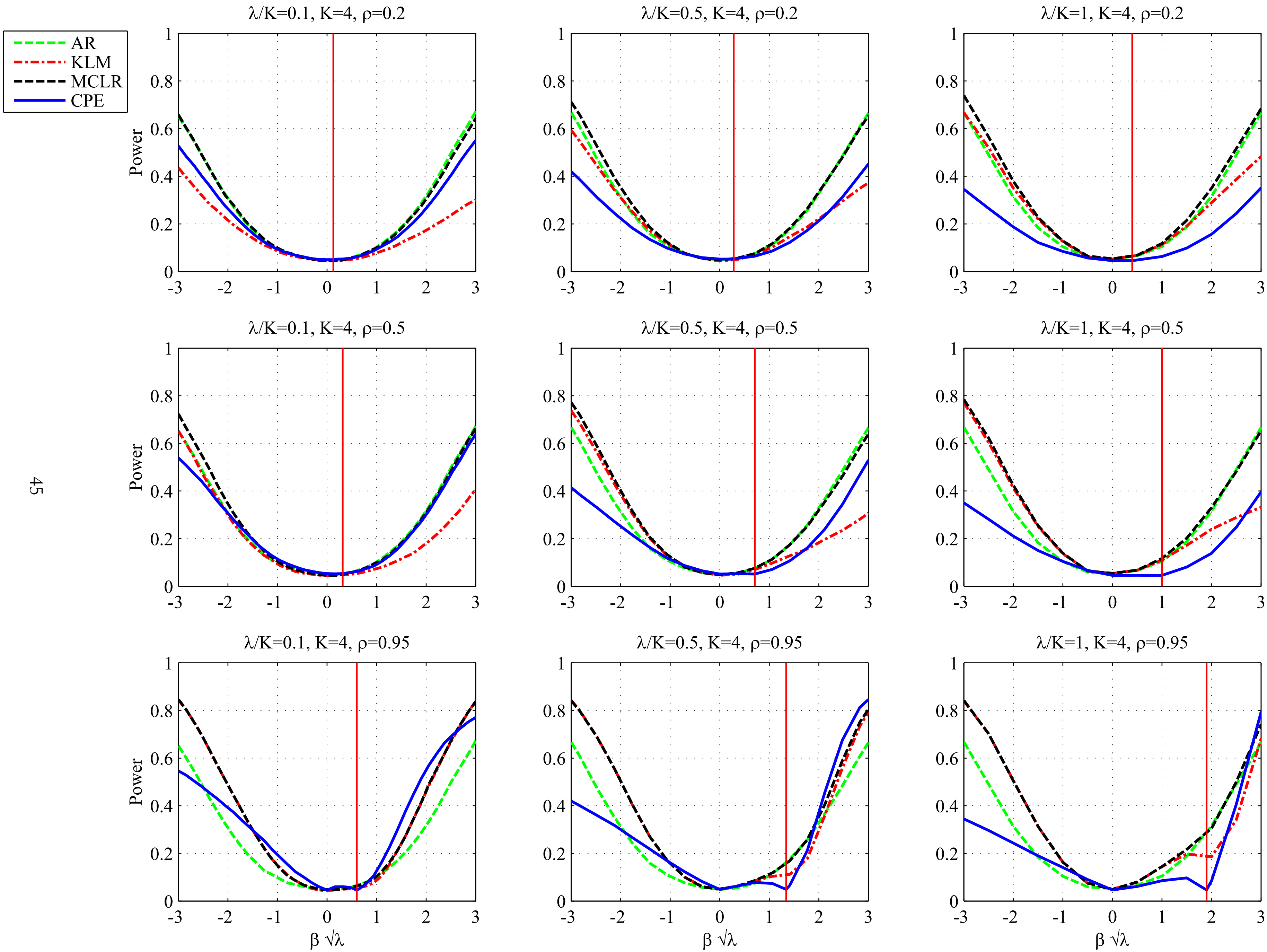

Figure 1B: Unconditional Power Plots, $k = 4$. Different values of λ/k and ρ.

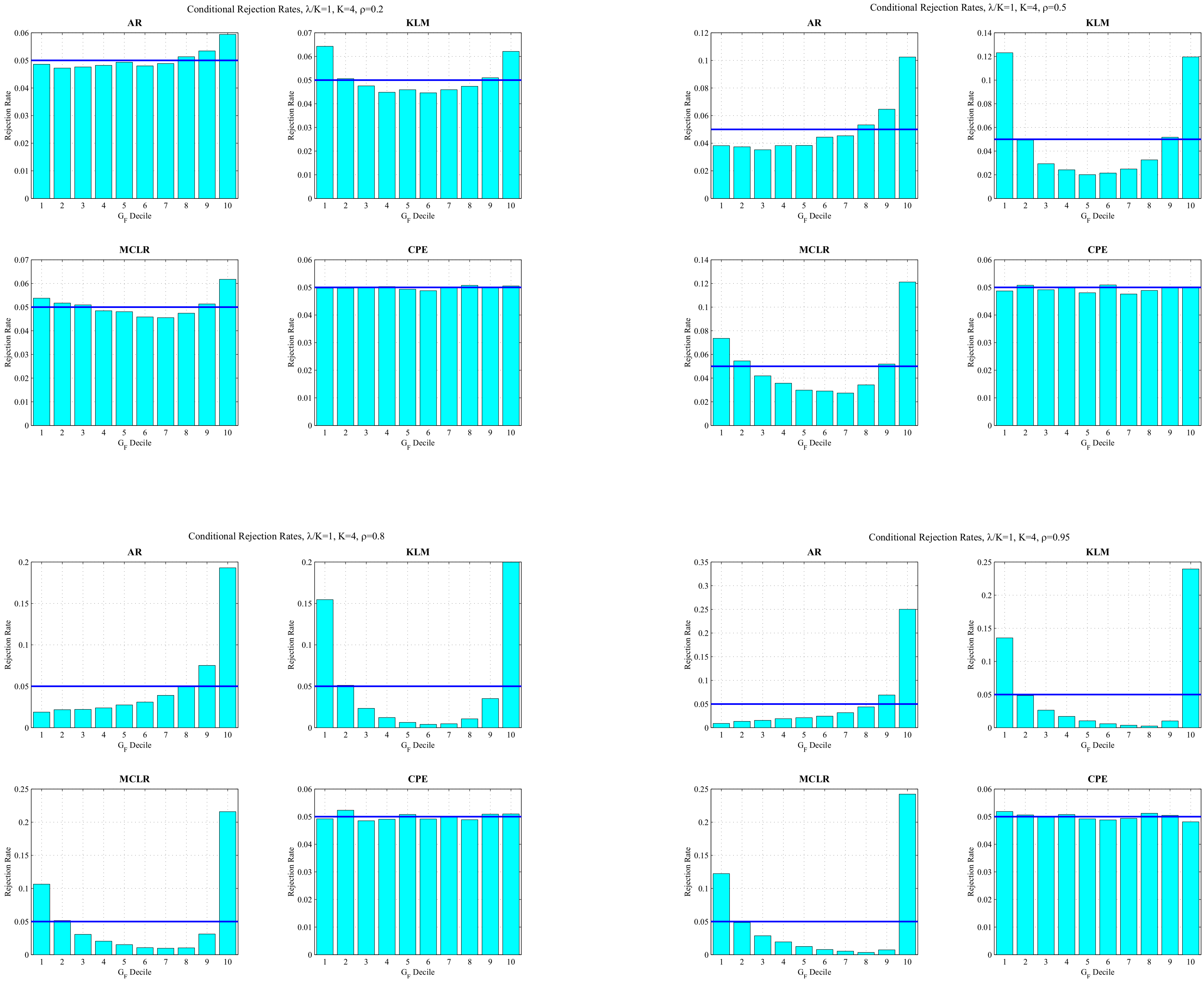

Figure 2A: Conditional Rejection Rates at the null ($\beta = 0$), for $k = 4$. G_F deciles on the x-axis.

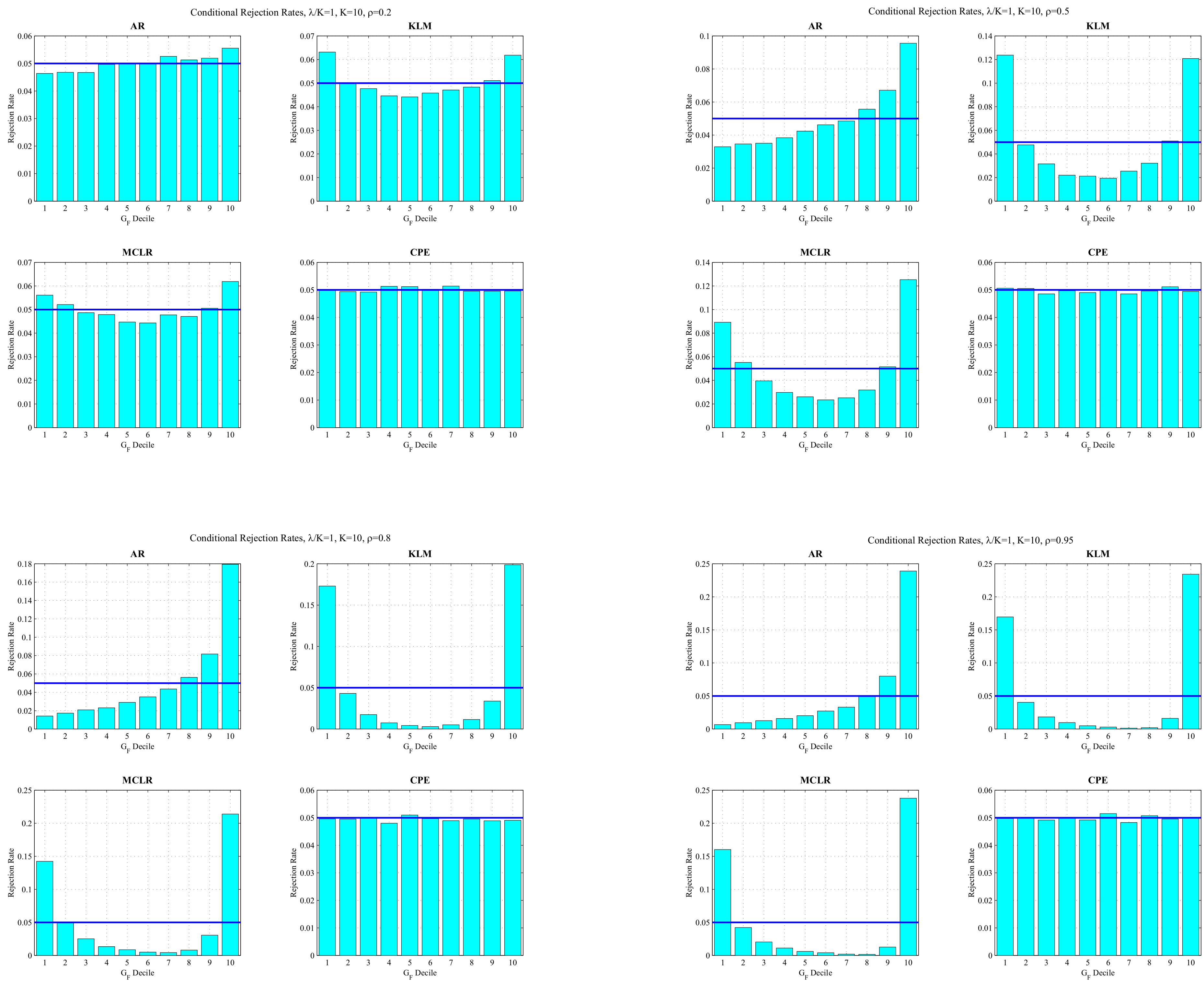

Figure 2B: Conditional Rejection Rates at the null ($\beta = 0$), for $k = 10$. G_F deciles on the x-axis.

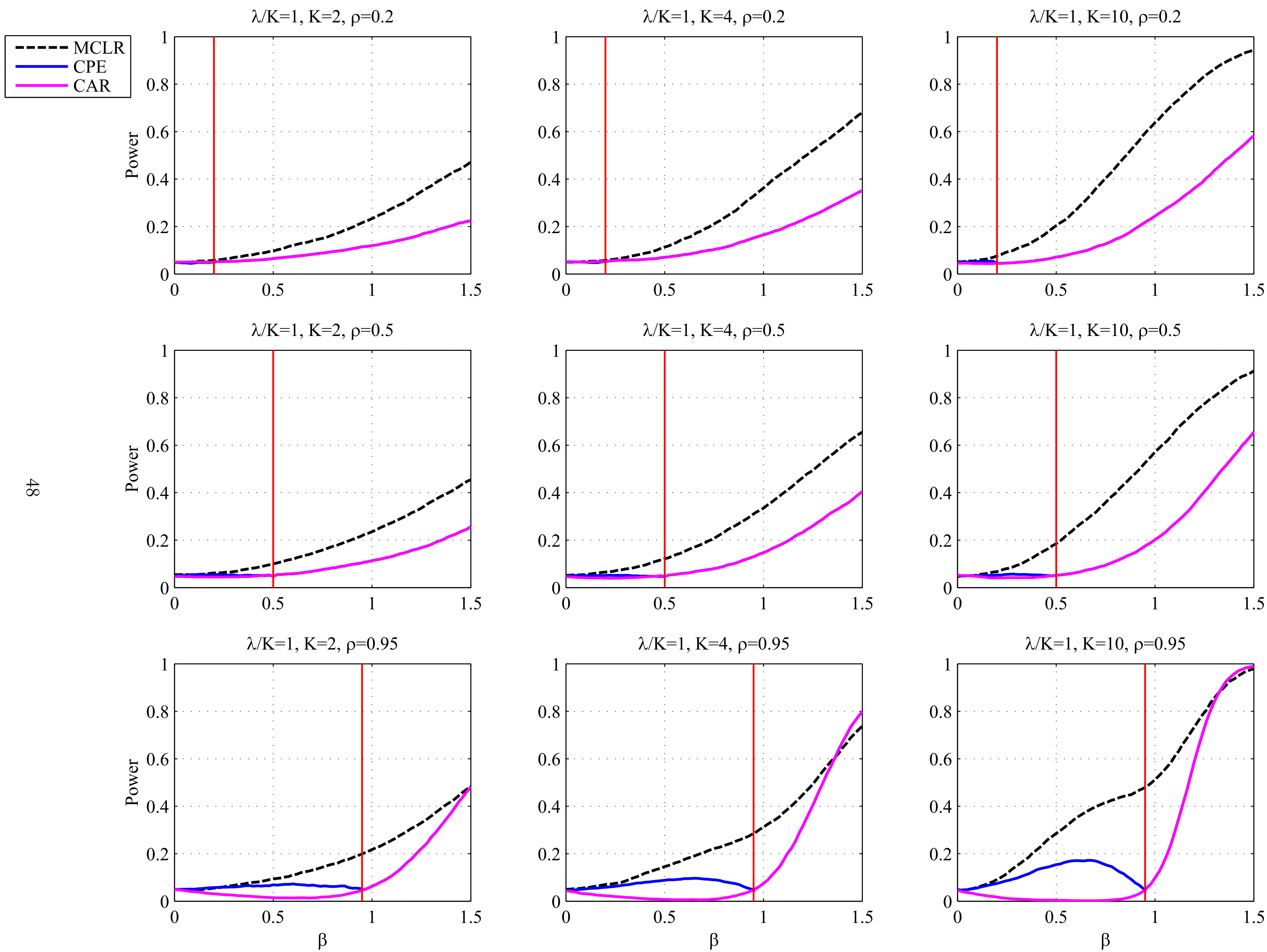

Figure 3A: Shortcoming of the Conditional Anderson-Rubin (CAR), different values of ρ and k.

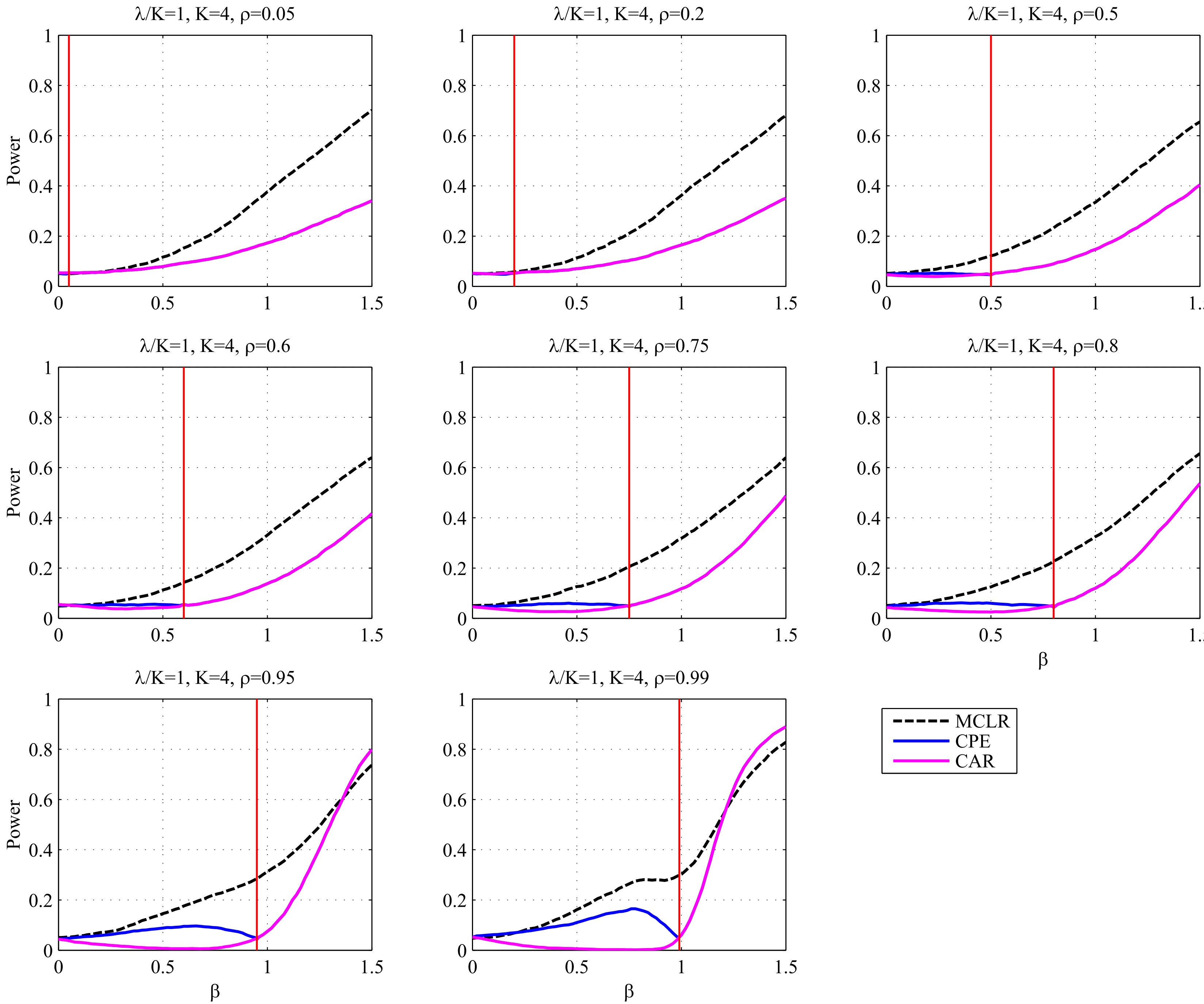

Figure 3B: Shortcoming of the Conditional Anderson-Rubin test (CAR), k = 4 (finer grid grid along the ρ dimension).

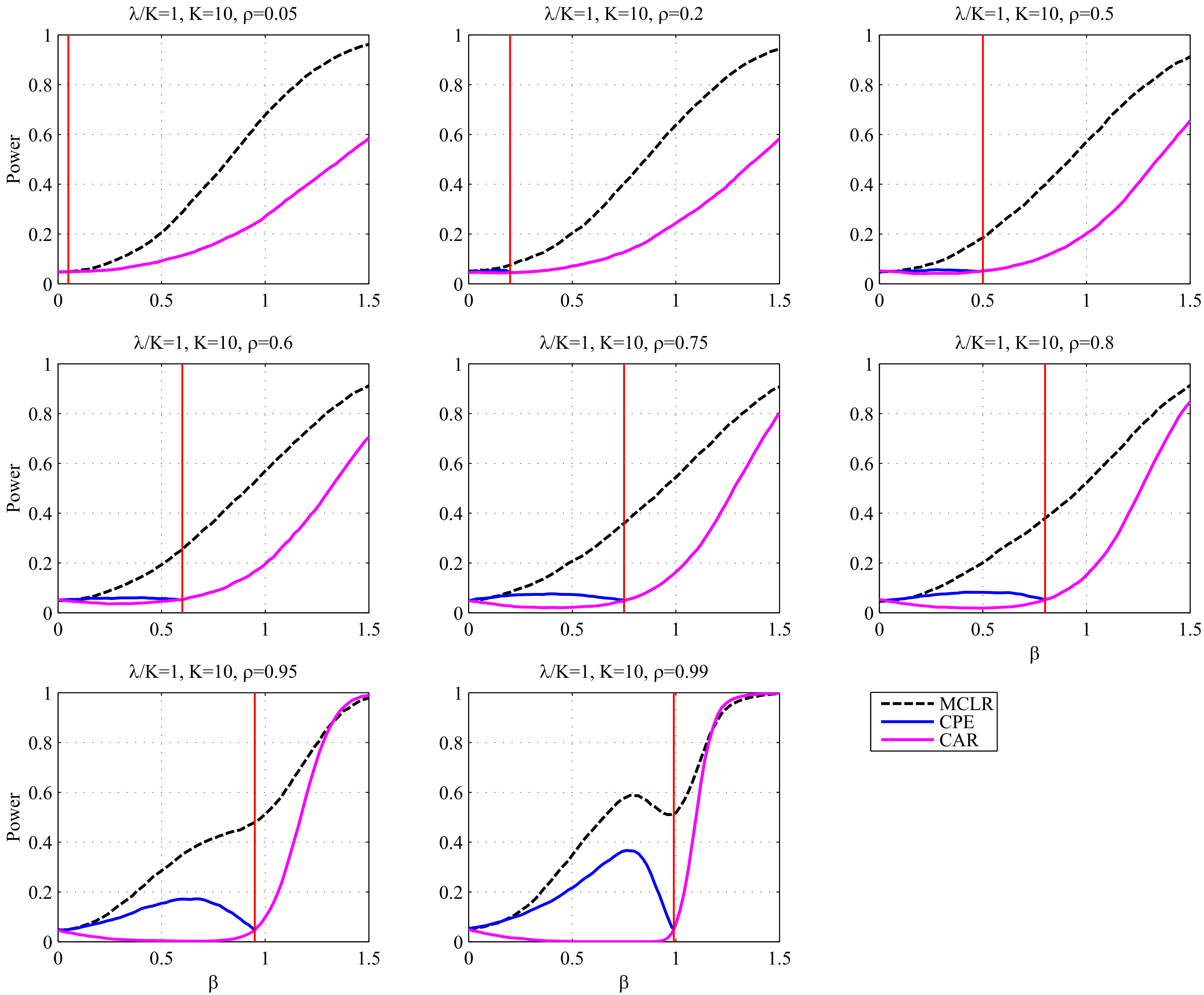

Figure 3C: Shortcoming of the Conditional Anderson-Rubin test (CAR), k = 10 (finer grid on the ρ dimension).

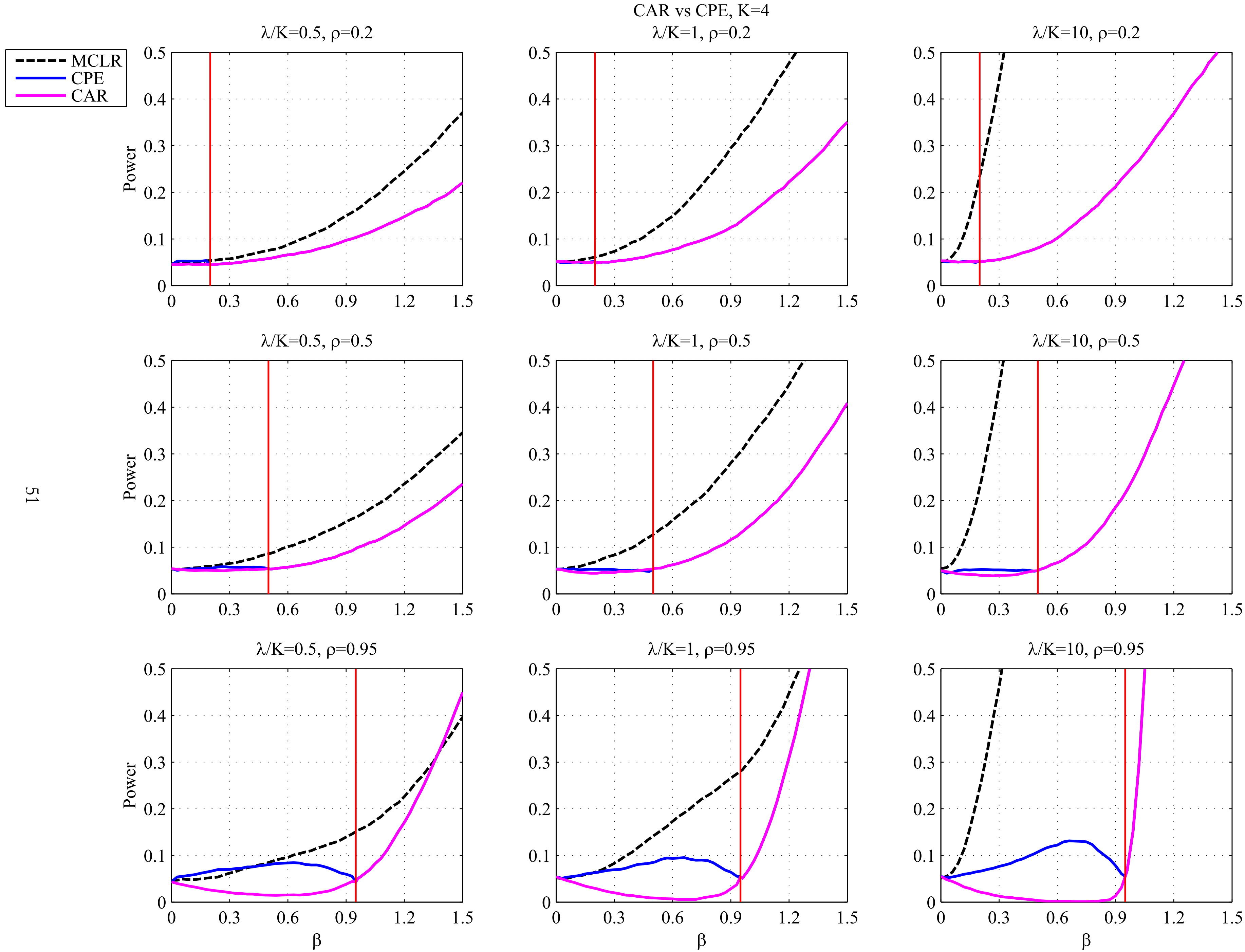

Figure 3D: Shortcoming of the Conditional Anderson-Rubin tests (CAR), $k = 4$. Different values of λ/k and ρ.

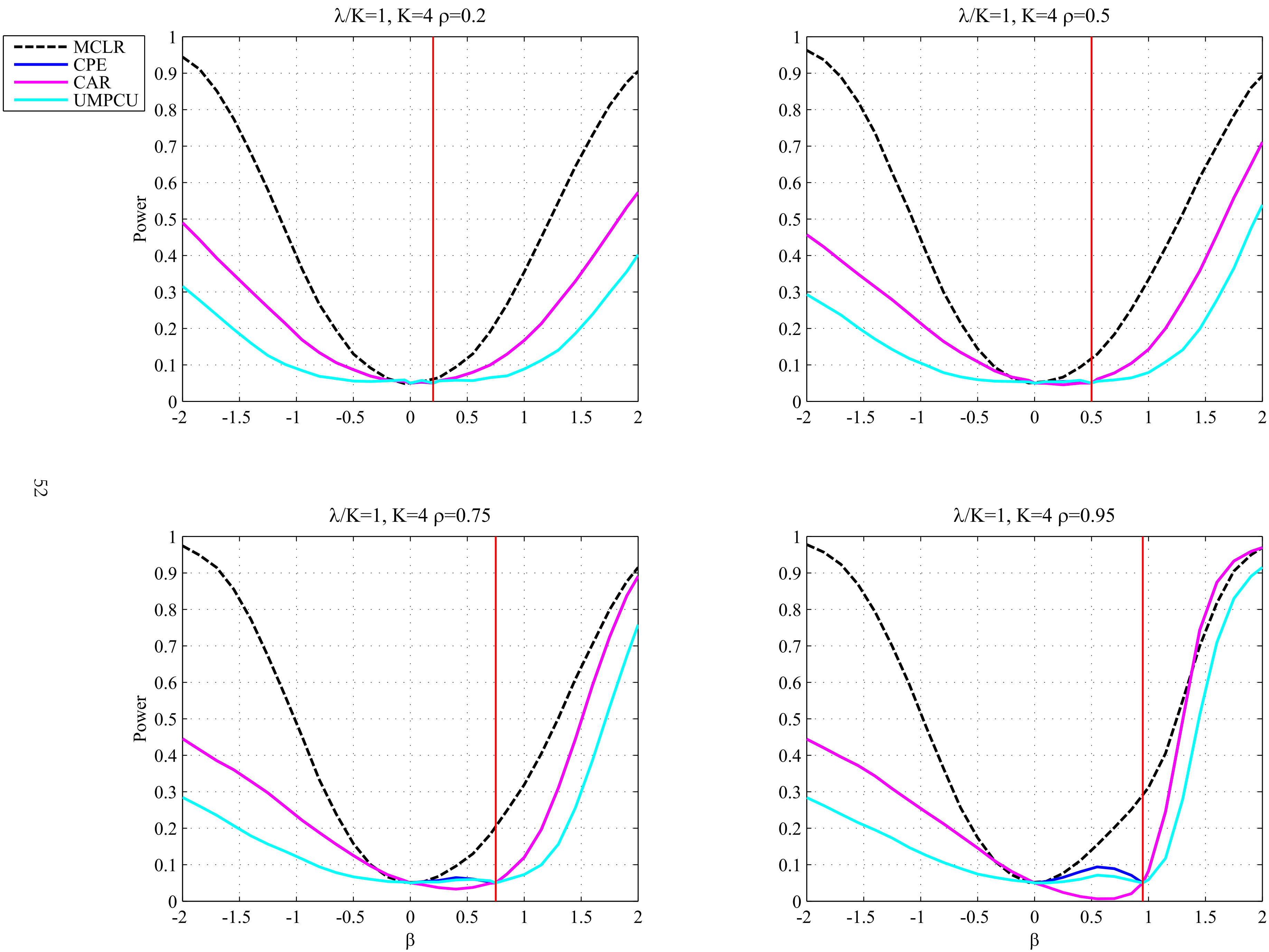

Figure 4A: Shortcoming of the UMPCU test, k = 4. Different values of ρ.

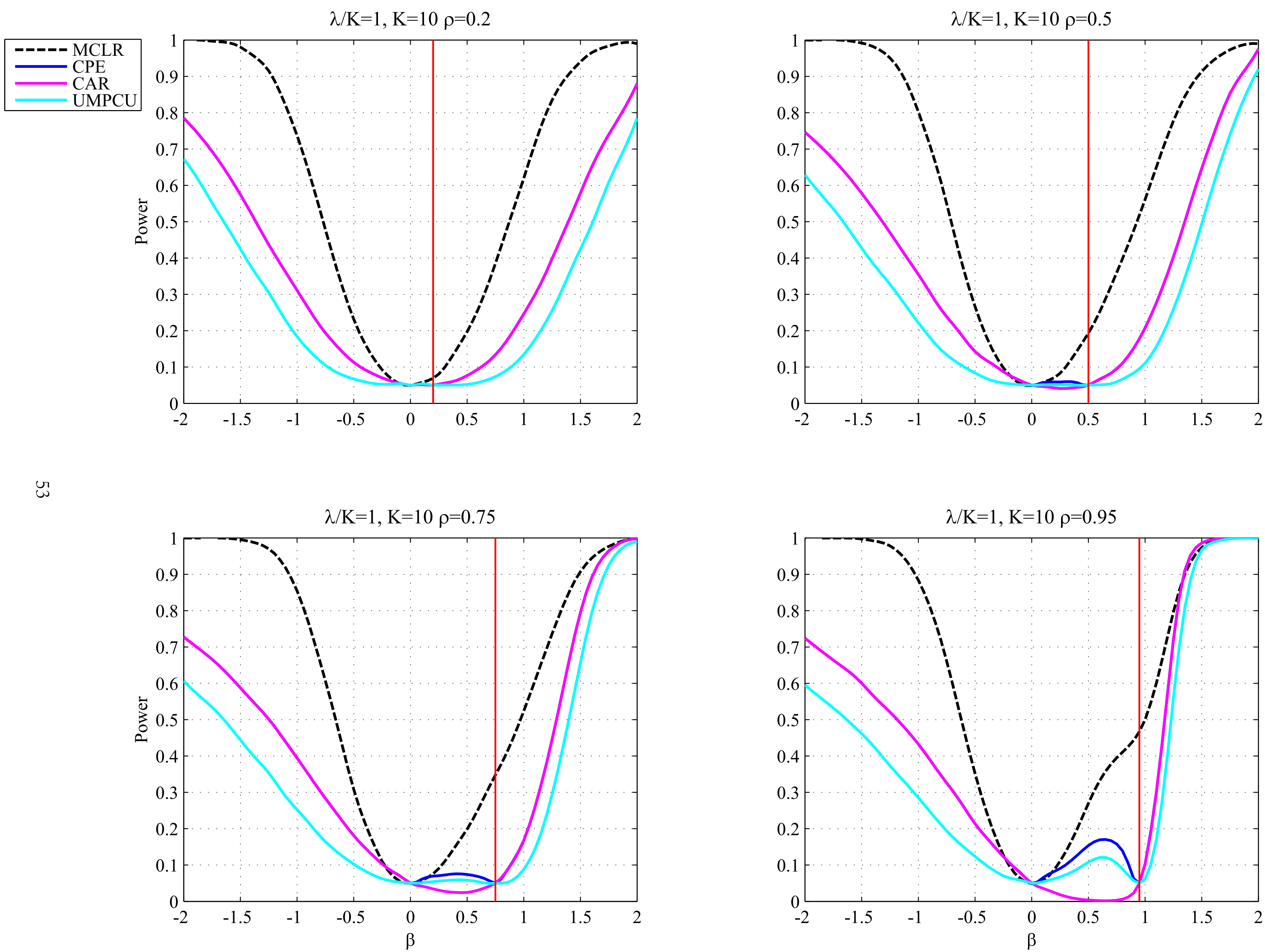

Figure 4B: Shortcoming of the UMPCU test, k = 10. Different values of ρ

References

Anderson, T. W., N. Kunitomo, and T. Sawa (1982): "Evaluation of the Distribution Function of the Limited Information Maximum Likelihood Estimator," *Econometrica*, 50, 1009–28.

Anderson, T. W., and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 20 (1), 46–63.

Andrews, D. W. K., M. J. Moreira, and J. H. Stock (2004): "Optimal Invariant Similar Tests for Instrumental Variables Regression," NBER Working Paper T0299.

Angrist, J., and B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings," *Quarterly Journal of Economics*, 106, 979–1014.

——— (2001): "Instrumental Variables and the Search for Identifications: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*, 15(4), 69–85.

Basu, D. (1964): "Recovery of Ancillary Information," *Sankhya*, 26, Series A, 3–16.

——— (1977): "On the Elimination of Nuisance Parameters," *Journal of the American Statistical Association*, 72(358), 355–366.

Becker, G. S. (1967): *Human Capital and the Personal Distribution of Income*. University of Michigan Press, Ann Arbor Michigan.

Bekker, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62(3), 657–681.

Bierens, H. (1982): "Consistent Models Specification Tests," *Journal of Econometrics*, 20, 105–34.

Billingsley, P. (1999): *Convergence of Probability Meausures*. Wiley, New York, second edn.

Birnbaum, A. (1962): "On the Foundation of Statistical Inference," *Journal of American Statistical Association*, 57, 269–306.

Blundell, R. W., and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, 655–679.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90(430), 443–50.

Card, D. (2001): "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69(5), 1127 – 1160.

Chamberlain, G. (2001): "Instrumental Variables Invariance and Minimax," unpublished manuscript, Department of Economics, Harvard University.

Chamberlain, G., and G. Imbens (2004): "Random Effects Estimators with Many Instrumental Variables," *Econometrica*, 72(1), 295–306.

Chao, J., and N. Swanson (2003): "Asymptotic Normality of Single-Equation Estimators for the Case with a Large Number of Weak Instruments," unpublished manuscript.

——— (2005): "Consistent Estimation With a Large Number of Weak Instruments," Econometrica, forthcoming.

Cox, D. R. (1958): "Some Problems Conected with Statistical Inference," *Annals of Mathematical Statistics*, 29, 357–72.

——— (1988): "Some Aspects of Conditional and Asymptotic Inference: A Review," *Sankhya*, 50, Series A, 314–337.

Cox, D. R., and D. V. Hinkley (1974): *Theoretical Statistics*. Chapman and Hall, London.

Donald, S. G., and W. K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69(5), 1161–1191.

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

Dufour, J. M. (1997): "Some Impossibility Theoremsin Econometrics with Applications to Structural and Dynamic Models," *Econometrica*, 65, 1365–1388.

——— (2003): "Identification, Weak Instruments, and Statistical Inference in Econometrics," *Canadian Journal of Economics*, 36, 767–808.

Dufour, J. M., and J. Jasiak (2001): "Finite Sample Limited Information Inference Methods for Structural Equations and Models with Generated Regressors," *International Economi Review*, 42, 815–843.

Dufour, J. M., and M. Taamouti (2005): "Projection Based Statistical Inference in Linear Structural Models with Possibly Weak Instruments," Econometrica, forthcoming.

Feigin, P. D. (1986): "Asymptotic Theory of Conditional Inference for Stochastic Processes," *Stochastic Processes and their Applications*, 22, 89–102.

Forchini, G., and G. Hillier (2003): "Conditional Inference for Possibly Unidentified Structural Equations," *Econometric Theory*, 19, 707–743.

——— (2004): "Ill-Posed Problems and Instruments' Weakness," unpublished manuscript.

Gleser, L., and J. Hwang (1987): "The Nonexistence of $100\,(1-\alpha)\,\%$ Confidence Sets of Finite Expected Diameter in Errors-in-Variables and Related Models," *Annals of Statistics*, 15, 1351–1362.

Hahn, J. (2002): "Optimal Inference with Many Instruments," *Econometric Theory*, 18, 140–168.

Hahn, J., and J. Hausman (2002): "A New Specification Test for the Validity of Instrumental Variables," *Econometrica*, 70(1), 163–189.

Hahn, J., J. Hausman, and G. Kuersteiner (2004): "Estimation with Weak Instruments: Accuracy of Higher-Order Bias and MSE Approximations," *Econometrics Journal*, 7, 272–306.

Hall, W. J., and R. M. Loynes (1977): "On the Concept of Contiguity," *The Annals of Probability*, 5(2), 278–282.

Hansen, C., J. Hausman, and W. K. Newey (2004): "Many Instruments, Weak Instruments, and Microeconomics Practice," unpublished manuscript, Department of Economics, MIT.

Hillier, G. (2004): "Hypothesis Testing in a Structural Equation: Similar and Invariant Tests," unpublished manuscript, University of Southampton.

Jansson, M., and M. J. Moreira (2004): "Optimal Inference in Regression Models with Nearly Integrated Regressors," unpublished manuscript.

Kalbfleisch, J. D. (1975): "Sufficiency and Conditionality," *Biometrika*, 62(2), 251–59.

Kalbfleisch, J. D. (1982): "Ancillary Statistics," in *Encyclopedia of Statistics*, ed. by K. S., and N. L. Johnson, vol. 1, pp. 77–81. Wiley.

Kleibergen, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781–1803.

Le Cam, L., and G. L. Yang (2000): *Asymptotics in Statistics: Some Basic Concepts*. Springer-Verlag, New York, second edn.

Lehmann, E. L. (1997): *Testing Statistical Hypotheses*. Springer, New York.

Lehmann, E. L., and F. W. Scholz (1992): "Ancillarity," in *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, ed. by M. Ghosh, and P. K. Pathak, vol. 17 of *Lecture Notes-Monograph Series*, pp. 32–51. Institute of Mathematical Statistics.

Moreira, M. J. (2001): "Tests with Correct Size When Instruments Can Be Arbitrarily Weak," unpublished manuscript, Department of Economics, University of California, Berkeley.

——— (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71(4), 1072–1048.

——— (2005): "Towards a General Theory of Hypothesis Testing for Structural Models," unpublished manuscript, Department of Economics, Harvard University.

Muirhead, R. J. (1982): *Aspects of Multivariate Statistical Theory*. Wiley, New York.

Nelson, C. R., and R. Startz (1990): "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator," *Econometrica*, 58, 967–976.

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67(3), 565–603.

Reid, N. (1995): "The Roles of Conditioning in Inference," *Statistical Science*, 10(2), 138–157.

Rothenberg, T. J. (1984): "Approximating the Distributions of Econometric Estimators and Test Statistics," in *Handbook of Econometrics, Volume II*, ed. by Z. Griliches, and M. D. Intriligator. Elsevier, New York: North Holland, pp. 881-935.

Severini, T. A. (1995): "Inference Based on Estimating Functions in the Presence of Nuisance Parameters: Comment," *Statistical Science*, 10(2), 187–189.

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

Startz, R., E. Zivot, and C. Nelson (2003): "Improved Inference in Weakly Identified Instrumental Variables Regression," unpublished manuscript, Department of Economics, University of Washington.

Stock, J. H., and M. Yogo (2002): "Testing for Weak Instruments in Linear IV Regression," NBER TWP 0284.

——— (2003): "Asymptotic Distributions of Instrumental Variables Statistics with Many Weak Instruments," unpublished manuscript.

Stock, J. H., M. Yogo, and J. H. Wright (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic Statistics*, 20, 518–529.

Sweeting, T. J. (1989): "On Conditional Weak Convergence," *Journal of Theoretical Probability*, 2, 461–474.

van der Vaart, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge UK.

——— (2002): "The Statistical Work of Lucien Le Cam," *The Annals of Statistics*, 30(3), 631–682.

Wang, J., and E. Zivot (1998): "Inference on a Structural Parameter in Instrumental Variables Regression with Weak Instruments," *Econometrica*, 66, 1389–1440.

Zivot, E., R. Startz, and C. R. Nelson (1998): "Valid Confidence Intervals and Inference in the Presence of Weak Instruments," *International Economic Review*, 39, 1119–1144.