Model Equivalence Tests in a Parametric Framework

Pascal Lavergne, Toulouse School of Economics

This version: April 2010

Preliminary draft – Comments welcome

Abstract

Empirical researchers commonly want to assess the validity of restrictions on parameters, appearing as an economic hypothesis, a consequence of economic theory, or an econometric modeling assumption. I propose a new theoretical framework to assess the *approximate* validity of multivariate restrictions in parametric models. I construct tests that are locally asymptotically maximin and locally asymptotically uniformly most powerful invariant. The tests are applied to three different empirical problems.

Keywords: Hypothesis testing, Parametric methods.

Address correspondence: Pascal Lavergne, Toulouse School of Economics-GREMAQ, Université Toulouse 1, 21 Allées de Brienne, 31000 Toulouse FRANCE. Email: pascal.lavergne@univ-tlse1.fr

If by the truth of Newtonian mechanics we mean that it is approximately true in some appropriate well defined sense we could obtain strong evidence that it is true; but if we mean by its truth that it is exactly true then it has already been refuted. I.J. Good (1981)

1 Introduction

A rather common objective in econometric or statistical modeling is to assess that some restrictions hold. For instance, practitioners often test whether a parametric model is correctly specified by embedding their model in one involving more parameters and testing for the significance of the extra coefficients. While in a test of significance, the researcher is typically hoping that the null hypothesis of insignificance will be rejected, in the case of a specification error test, the researcher often hopes the null will be *accepted*. Specification testing is by no means an atypical situation, and there are many instances where we would like to obtain evidence in favor of restrictions that appear as (i) an economic hypothesis, for instance constant returns to scale in an aggregate production function; (ii) a consequence of economic theory, for instance homogeneity of demand in prices and income as implied by consumer rationality; (iii) a key assumption to estimate a structural model, such as exogeneity.

It has been early acknowledged that applied researchers are often looking for evidence *in favor* of a particular hypothesis. For instance, Berkson (1942) argues that the P-value of a significance test can be used as evidential measure in favor of the null hypothesis. This however has been strongly criticized, see e.g. the discussion of Berger and Sellke (1987) and the references cited therein. Some argue in favor of Bayes factors as introduced by Jeffreys (1961), see Kass and Raftery (1995) and the references therein. Others, such as Good (1983, 1992), advocates for a compromise of Bayesian and non Bayesian approaches. Andrews (1994) points out under certain asymptotics there exists a correspondence between P-values and Bayesian posterior odds.

The goal of this work is to develop a testing procedure for assessing the *approximate*

validity of restrictions in parametric models. The interest of approximate hypotheses has been long recognized in statistics, see e.g. Hodges and Lehmann (1954). Leamer (1988) argues that "genuinely interesting hypotheses are neighborhoods, not points. No parameter is exactly equal to zero; many may be so close that we can act as if they were zero," see also Good (1981) in statistics or McCloskey (2001) in economics among others. We consider the approximate validity of the restrictions of interest as the *alternative* hypothesis to reflect where the burden of proof is placed. This is known in biostatistics as *equivalence testing*, where the practitioner wants to assess that a parameter, usually the difference in bio-effect between two formulations of the same molecule, is close to zero, see Lehman and Romano (2005), Wellek's (2003) monograph, and Senn's (2001) review. Another application of this principle is provided by Dette and Munk (1998) for specification testing. Finally, our approximate alternative hypothesis concentrates around the sharp restrictions of interest as the sample size increases to formalize that we are interested in showing that our restriction is as close to be fulfilled as made possible by our data, see Rosenblatt (1962) for an early example. An approach closely related to ours is proposed by Romano (2005) for equivalence testing, see also Borovkov (1998) for related results. For testing an univariate restriction on parameters of the form $g(\theta) = 0$, Romano considers

$$H_n : \ |g(\theta)| \geq \delta/\sqrt{n} \qquad \text{against} \qquad K_n : \ |g(\theta)| < \delta/\sqrt{n} \ .$$

The alternative hypothesis of interest K_n is then a neighborhood of the hypothesis of interest that becomes narrower as sample size, and thus information, increases. The related test yields a decision of whether *a set of parameter values* that are close to the restrictions is *consistent* with the data at hand.

By contrast to the latter approach, our framework does not directly focus on the restricted parameters themselves, but on the consequences of imposing these restrictions. Following Akaike (1973) and Vuong (1989), among others, we focus on the effect of the restrictions as measured by the Kullback Leibler Information Criterion (KLIC), which is a natural discrepancy measure between the unrestricted model and the restricted one.

Hence the alternative hypothesis of interest states that the KLIC is less than a small quantity that decreases when the sample size increases. This allows in particular to consider multivariate restrictions on parameters, which has not been dealt with in previous work. For our approximate shrinking alternative hypothesis, we derive a test based on the usual likelihood-ratio (LR) statistic, but that uses a decision rule different from that of a significance test: the alternative hypothesis is accepted for small values of the statistic, and the critical value is not derived under the assumption that the restrictions perfectly hold. We label this approach *model equivalence testing*.

While our main model equivalence test is based on the LR statistic, it seems natural to investigate whether equivalent procedures can be derived, as is the case in significance testing. We show indeed that one can consider different asymptotically equivalent formulations of the testing problem. The first one relies on a Hausman-Wald approach, following the terminology of Gourieroux and Monfort (1989), and evaluates how the restrictions affect the *whole parameter vector*. The second relies on a score approach and evaluates whether the expected score of the restricted model is close to zero. The third relies on a Wald approach and is similar to Romano's equivalence test in the case of univariate restrictions. To each set of hypotheses correspond a different test. We show that the four tests are locally asymptotically maximin and locally most powerful against the hypothesis that the restrictions perfectly hold. They are also locally asymptotically most powerful in the class of tests invariant to orthogonal transformations of the parameter. For testing an univariate restriction, the proposed tests are equivalent to the tests proposed by Romano (2005), and thus are locally asymptotically uniformly most powerful.

One may wonder whether and why a new approach is needed. The pervasive observation that practitioners commonly use significance tests when they actually intend to accept the insignificance hypothesis should be enough motivation for a new look at this issue. A significance test entertained at usual nominal levels can never accept the null and thus cannot assess the validity of restrictions, even in an approximate sense. This is because it tunes the odds of falsely rejecting a null hypothesis, but does not control the probability of *falsely not rejecting it*. Do confidence intervals or region could provide

such information? These are defined as sets of parameters values that cannot be rejected by a significance test, so they do not provide a suitable answer either. Within the significance testing approach, Andrews (1989) proposes approximations of the asymptotic inverse power function as an aid to interpret non significant outcomes. Andrews' inverse power approximations are based on the Wald test and are thus not invariant to nonlinear transformations of restrictions under scrutiny, see e.g. Gregory and Veall (1985). More crucially, several issues surround power calculations, as summarized by Hoenig and Heisey (2001).[1] Finally, *evaluating* the asymptotic power of a significance test of given level does not directly provide positive evidence in favor of the restrictions under consideration. Our proposal instead can be viewed as *controlling* the asymptotic power of a significance test for a set of parameters values, so as to obtain reliable evidence in favor of the approximate restrictions.

The paper is organized as follows. In Section 2, I setup the testing framework based on the KLIC, I derive the model equivalence LR test, as well as alternative formulations and tests. In Section 3, I study the local asymptotic properties of the tests. In Section 4, I discuss implementation of these tests through three examples. In Section 5, I conclude by suggesting directions for extensions and future research.

2 Testing Framework and Procedures

Let us introduce the basic setup considered throughout this paper. To focus on the main issues, we deal with unconditional models, but all our results can be extended to conditional models under standard assumptions, such as a fixed or i.i.d. design of the conditioning variables. We observe a random sample $\{X_t, t = 1, \dots n\}$ from X, whose probability density $f(\cdot, \theta_0)$ belongs to a parametric family of densities $\{f(\cdot, \theta) : \theta \in \Theta\}$.

[1] In some applied sciences where they are common practice, the debate surrounding post-experiment power calculation is quite vivid and seems to be an old one: in his 1958 book, Cox writes "Power is important in choosing between alternative methods of analyzing data and in deciding on an appropriate size of experiments. It is quite irrelevant in the actual analysis of data."

Denote by $\mathbb{E}_{\theta_0}$ the expectation when θ_0 is the parameter value. We are interested in assessing the validity of some multivariate restrictions on parameters of the form $g(\theta_0) = 0$, where $g(\cdot)$ is a function from $\mathbb{R}^p$ to $\mathbb{R}^r$, $1 \leq r < p$. Let

$$\theta_0^c = \arg \max_{\theta \in \Theta, g(\theta)=0} \mathbb{E}_{\theta_0} \log f(X, \theta) \,. \tag{2.1}$$

be the pseudo-true value of the maximum-likelihood estimator under the constraint, see e.g. Sawa (1978), and note that θ_0^c depends on θ_0 only. Denote by ∇_θ differentiation with respect to θ, and by $\nabla_{\theta,\theta'}$ second differentiation. We make the following assumptions.

Assumption A *(a) The densities $f(X, \theta)$, $\theta \in \Theta$ are defined with respect to a common dominating measure ν. (b) The set Θ is an open bounded subspace of $\mathbb{R}^p$. (c) $f(\cdot, \theta_1) \equiv f(\cdot, \theta_2)$ implies $\theta_1 = \theta_2$. (d) The densities $\sqrt{f(\cdot, \theta)}$ are continuously differentiable in θ almost everywhere. (e) The function $l(\cdot, \theta) = \log f(\cdot, \theta)$ is twice continuously differentiable in θ almost everywhere. There exists a function $\bar{l}(x)$ such that $\|\nabla^2_{\theta,\theta'} l(x, \theta)\| < \bar{l}(x)$ and $\mathbb{E}_\theta \bar{l}(X) < \infty$ uniformly over a neighborhood of θ_0. (f) $I(\theta) \equiv \mathbb{E}_\theta \left[\nabla_\theta l(X, \theta) \nabla'_\theta l(X, \theta)\right]$ exists, is continuous and positive definite uniformly over a neighborhood of θ_0.*

Assumption B *(i) $g(\cdot)$ is continuously differentiable and $\nabla_\theta g(\cdot)$ is of full rank r uniformly over a neighborhood of θ_0. (ii) θ_0^c is unique.*

2.1 KLIC-Based Testing

Following Akaike (1973, 1974), Sawa (1978), and Vuong (1989), among others, we consider as a measure of closeness of a model to the true distribution the Kullback-Leibler Information Criterion defined as

$$KLIC = \mathbb{E}_{\theta_0} \left[\log \frac{f(X, \theta_0)}{f(X, \theta_0^c)} \right] \,.$$

This measure is always positive and zero if and only if the restrictions perfectly hold, see Vuong (1989). We consider as the hypothesis to assess

$$K_n^{LR} : \; 2 \; KLIC < \delta^2 / n \,.$$

This means that imposing the constraint does not affect the expected likelihood by more than $\delta^2/2n$. Our null hypothesis is the complement of the alternative, that is

$$H_n^{LR} : \ 2\ KLIC \geq \delta^2/n\,.$$

Though our framework is reminiscent of local power analysis of significance tests, it is only our inability to confirm a sharp hypothesis, together with our will to be as tight as possible around this hypothesis, that motivates this formulation. One should also note that our setup is different from the one envisaged in model selection, where one aims to choose the unrestricted model if $KLIC > 0$ and the restricted one if $KLIC = 0$. In that aim, the penalty term added to the likelihood-ratio statistic, which estimates the KLIC, is used only to ensure that the correct and most parsimonious model is chosen asymptotically, see e.g. Sin and White (1996) for general results on this approach.

Our shrinking hypothesis setup with threshold δ^2/n puts us in the most difficult but manageable situation. Would the threshold go towards zero faster than $n^{-1/2}$, all distributions in K_n^{LR} would be contiguous to some distributions in H_n^{LR} and then would not be distinguishable from H_n^{LR}. One may want to consider a fixed alternative hypothesis instead of a shrinking one, i.e. set the limit at Δ^2 instead of δ^2/n. From a practical viewpoint, the choice of Δ^2 in a fixed hypothesis setup would become even more central to the procedure, and it is unlikely that practitioners could reach a consensus on which value to consider in specific applications. Moreover, considering a more restrictive hypothesis as the sample size increases formalizes that our ultimate goal is to assess the validity of some restrictions, so it makes sense to adopt an asymptotic setup that explicitly acknowledges our aim. In addition, if one were to devise a good test for a fixed hypothesis but uses in practice a sample size dependent hypothesis, the resulting test can have very low power, see Romano (2005) for an example.

Consider the (quasi-)maximum likelihood (ML) estimators of θ_0 and θ_0^c

$$\widehat{\theta}_n = \arg\sup_{\Theta} L_n(\theta) = \arg\sup_{\Theta} \sum_{t=1}^{n} l(X_t, \theta) \quad \text{and} \quad \widehat{\theta}_n^c = \arg \sup_{\Theta, g(\theta)=0} L_n(\theta)\,.$$

The likelihood-ratio (LR) statistic is $2\ LR_n = 2\left[L_n(\widehat{\theta}_n^c) - L_n(\widehat{\theta}_n)\right]$. The LR model equivalence of H_n^{LR} against K_n^{LR} is defined as $\pi_n^{LR} = \mathbb{I}\left[2\ LR_n \leq c_{\alpha,r,\delta^2}\right]$, where $\Pr\left[\chi_r^2(\delta^2) \leq c_{\alpha,r,\delta^2}\right] = \alpha$, that is, the critical value is the α quantile of a noncentral chi-square distribution with r degrees of freedom and non centrality parameter δ^2. This stands in contrast to the critical value of a significance test, which is the $1-\alpha$ quantile of a central chi-square distribution.

The choice of δ^2 is key because it defines the hypothesis under test. Clearly, the smaller δ^2, the more stringent the equivalence hypothesis. If one accepts the equivalence hypothesis for a particular value δ^2, then the outcome will be unchanged for any model equivalence $2\ KLIC \leq \gamma^2/n$ with $\gamma^2 > \delta^2$. While its specific value should be tailored to the specific application at hand, some general guidelines can be offered. From the form of the equivalence hypothesis, one can interpret δ^2 as the maximum value at which we are ready to declare model equivalence, that is the equivalence hypothesis for $n = 1$. This allows us to choose δ^2 independently of the sample size at hand. Now it may appear uneasy to select δ^2 on these grounds because the KLIC of a model has no natural upper bound, but when the distribution of X is discrete. In any application however, it is often easy to determine such an upper bound by considering a model that has already been judged, on other grounds, non-equivalent to the unrestricted model. Take for instance the case of a standard normal regression model then a rough lower bound for $\mathbb{E}_{\theta_0} \log f(X,\theta)$ is the expected log-likelihood for a base model with no explanatory variable. Hence we can select δ^2 as a fraction of what has been gained by adding explanatory variables to the base model.[2] While this quantity is not known, it can be consistently estimated from the sample. Formally, using such sample information will not affect the asymptotic properties of the test. This possibility is illustrated later on in some of our applications in Section 4.

[2]This precludes application to the case where the restrictions of interest are that all the parameters but the constant are zero. However, I cannot think of an application where it would be of interest to assess that such restrictions hold. Testing that these do not hold is clearly useful and can be done through a significance test.

2.2 Alternative Formulations

As is well known, the validity of the restrictions $g(\theta_0) = 0$ can be formulated in different ways, and these different formulations yield different significance tests. We now show that there also exist different asymptotically equivalent formulations of our approximate hypothesis K_n^{LR}.

Lemma 2.1 *Under Assumptions A and B, if* $2\ KLIC = O(n^{-1})$*, then* $2\ KLIC$ *is equal to any of*

$$(\theta_0 - \theta_0^c)' I(\theta_0) (\theta_0 - \theta_0^c) (1 + o(1)) , \tag{2.2}$$

$$\mathbb{E}_0 \nabla_\theta' l(X, \theta_0^c) I^{-1}(\theta_0) \mathbb{E}_0 \nabla_\theta l(X, \theta_0^c) (1 + o(1)) \tag{2.3}$$

$$g'(\theta_0) \left[\nabla_\theta' g(\theta_0) I^{-1}(\theta_0) \nabla_\theta g(\theta_0)\right]^{-1} g(\theta_0) (1 + o(1)) . \tag{2.4}$$

The three asymptotic approximations of $2\ KLIC$ in Lemma 2.1 yield three alternative formulations of the testing problem as well as the corresponding tests. The Hausman-Wald approach considers the hypotheses

$$H_n^H : \ (\theta_0 - \theta_0^c)' I(\theta_0) (\theta_0 - \theta_0^c) \geq \delta^2/n \quad \text{against} \quad K_n^H : \ (\theta_0 - \theta_0^c)' I(\theta_0) (\theta_0 - \theta_0^c) < \delta^2/n .$$

The alternative hypothesis here involves a norm of the difference between the true and pseudo-true values. This norm is defined defined through the information contained in the model. Such a standardization amount to a change of units and make the different components comparable, which is useful when considering parameters with possibly different units: even in a standard linear regression, the parameter vector includes the intercept, the different slopes, and the error's variance. In considering a t-test about a mean, Arrow (1960) consider that the "economically significant difference" should be measured in standard deviations units. It is therefore interesting to note that such a standardization (through the asymptotic variance of the ML estimator) appears naturally in our approach. This formulation also suggest that δ^2 should be relatively small. In our applications of Section 4, sensible values for δ^2 appear to be within a pretty tight range.

Finally, the above formulation makes clear that the model equivalence hypothesis allows for local misspecification of the restricted model.

The score approach is based on

$$H_n^S : \ \mathbb{E}_0 \nabla_\theta' l(X, \theta_0^c) I^{-1}(\theta_0) \mathbb{E}_0 \nabla_\theta l(X, \theta_0^c) \geq \delta^2/n$$

$$\text{against} \quad K_n^S : \ \mathbb{E}_0 \nabla_\theta' l(X, \theta_0^c) I^{-1}(\theta_0) \mathbb{E}_0 \nabla_\theta l(X, \theta_0^c) < \delta^2/n \ .$$

The alternative of interest thus focuses on whether the expected score vector of the restricted model is close to zero in the metric defined by $I^{-1}(\theta_0)$. Finally, the Wald approach considers

$$H_n^W : \ g'(\theta_0) \left[\nabla_\theta' g(\theta_0) I^{-1}(\theta_0) \nabla_\theta g(\theta_0) \right]^{-1} g(\theta_0) \geq \delta^2/n$$

$$\text{against} \quad K_n^W : \ g'(\theta_0) \left[\nabla_\theta' g(\theta_0) I^{-1}(\theta_0) \nabla_\theta g(\theta_0) \right]^{-1} g(\theta_0) < \delta^2/n \ .$$

Here the hypotheses focus on the restrictions themselves and have a clear intuitive content. In particular, it provide further insight on the choice of δ^2. Relying on this formulation, we can interpret the model equivalence hypothesis as a region of the parameters values "centered" around the restrictions of interest. Take for example a univariate restriction of the form $\theta_{01} = 0$. Model equivalence is then declared if the parameter belongs to $(-\delta\sigma_{01}/\sqrt{n} \ ; \delta\sigma_{01}/\sqrt{n} \ )$, where $\sigma_{01}/\sqrt{n}$ is the standard deviation of $\widehat{\theta}_{01}$. That is the value θ_{01} should be within δ standard deviations of zero. For interpretation's sake, this interval can be approximated by using the standard error of $\widehat{\theta}_{01}$. The Wald formulation thus allows to recast the equivalence hypothesis in terms of parameter values and the inherent variability in estimation. Such an *equivalence interval* can guide our choice and ease the interpretation of the test's results by providing direct information on the set of parameter values that is accepted by the test. The intuition extends to an equivalence hypothesis about a multidimensional parameter when testing a multidimensional hypothesis of the form $\theta_{01} = \mathbf{0}$.

To each alternative sets of hypotheses corresponds a different model equivalence test. Define the Hausman-Wald, score, and Wald statistic, respectively as

$$\begin{aligned} H_n &= n\left(\widehat{\theta}_n - \widehat{\theta}_n^c\right)' I(\widehat{\theta}_n)\left(\widehat{\theta}_n - \widehat{\theta}_n^c\right) \\ S_n &= n\nabla_\theta' L_n(\widehat{\theta}_n^c) I^{-1}(\widehat{\theta}_n)\nabla_\theta L_n(\widehat{\theta}_n^c) \\ W_n &= n g'(\widehat{\theta}_n)\left[\nabla_\theta' g(\widehat{\theta}_n) I^{-1}(\widehat{\theta}_n)\nabla_\theta g(\widehat{\theta}_n)\right] g(\widehat{\theta}_n)\,. \end{aligned}$$

Then each test π_n^J, $J = H, S, W$ is similarly defined as $\pi_n^J = \mathbb{I}\left[J_n \leq c_{\alpha,r,\delta^2}\right]$. Alternatively, the information matrix could be approximated by

$$I_n(\theta) = n^{-1}\sum_{t=1}^{n} \nabla_\theta \ln f(X_t;\theta)\nabla_\theta' \ln f(X_t;\theta) \qquad \text{or} \qquad -n^{-1}\sum_{t=1}^{n} \nabla^2_{\theta\theta'} \ln f(X_t;\theta)\,,$$

without altering the asymptotic properties of each test, which is useful for conditional models. However, it is well known that use of the last formula can yield non positive definite estimates, so this should not be recommended in practice.

All alternative formulations involve the whole parameter vector in general, even when the restrictions concern only a subset of them. All tests are invariant to linear transformations of the parameter space, but only the LR test is invariant to nonlinear reparametrization. The LR and Hausman-Wald are invariant to nonlinear transformation of the restrictions, as well as the score test that uses the outer product of gradient estimator of the information matrix, while the Wald test is invariant to linear transformations only.

2.3 Critical Values

While critical values are non-standard, they can be readily obtained from most statistical softwares. Tables 1 to 6 gives critical values for the test at 10% and 5% for δ^2 varying from 0.1 to 10 and r =1 to 6. It is seen that the critical values increase at a much slower rate than δ^2 and are generally below δ^2 (except for low values of δ^2). Figure 1 depicts the asymptotic power curves of the test for values of r, α, and δ^2, selected to illustrate their influence on the tests' power. It is seen that the power is always maximum when $KLIC = 0$, that is when the restrictions perfectly hold, but never attains one.

However, as we will show, no test can achieve a larger power at zero. In nature, the test is "tough" with the restrictions to be assessed. This is the price that we pay for controlling the probability of falsely confirming an hypothesis that narrows with the sample size. The power is increasing in, and pretty sensitive to, δ^2 and α. As previously discussed, the choice of δ^2 determines the limit of model equivalence and should be tailored to the particular application, see also Section 4. The level corresponds to the probability of falsely accepting model equivalence. It might also be tuned to obtain a test with reasonable power at zero, keeping in mind that the smaller the level the higher evidence we obtain in favor of model equivalence.

Since the test statistic is the same as in a significance test, we can interpret a model equivalence test as a significance test in reverse that controls the power for some values of the parameters space, as recommended for instance by Lehmann (1958) and Arrow (1960). If $\Lambda(\cdot)$ is the power function of the model equivalence test, then $1 - \Lambda(\cdot)$ is the power function of a significance test that tests $g(\theta_0) = 0$ for which the level is chosen so that the power has some predetermined value when $2KLIC = \delta^2/n$. In this sense, it is the inverse problem that of the one analyzed by Andrews (1989), who considers a significance test of given level and determines at which points of the parameter space the test has some predetermined power. Alternatively, we can interpret the non-significant outcome of a significance test by looking at the corresponding model equivalence test. For instance, if a significance test does not reject the sharp hypothesis $g(\theta_0) = 0$ at 5% level, this corresponds to accepting model equivalence at 5% level with $\delta^2 = 13$ for $r = 1$, 15.44 for $r = 2$, or 18.57, for $r = 4$ respectively. As argued above and illustrated below by our examples, such values appear to be extremely liberal.

3 Practical Applications

We here illustrate how our tests can be implemented in specific applications. In so doing, we also exemplify how our general guidelines for the choice of δ^2 can be put in action.

Example 1: Testing an Economic Hypothesis. We consider here a cross-country regression in the spirit of Mankiw and al. (1992), using pooled data on 86 countries averaged over the 1960s, 1970s and 1980s from King and Levine (1986), as analyzed by Stengos and Liu (1999). Explanatory variables include GDP60, the 1960 level of GDP; POP, population growth (to which 0.05 is added to account for depreciation rate and technological change); SEC, the enrollment rate in secondary schools; INV, the share of output allocated to investment; two dummy variables D70 and D80, acting as fixed effects for the seventies and the eighties. The model is estimated by OLS and yields

$$\begin{array}{lllll} \text{Growth} = & 0.0299 & -\ 0.0117\ D70 & -\ 0.0300\ D80 & +\ 0.0286 \log(INV) \\ & (0.0285) & (0.0032) & (0.0033) & (0.0041) \\ & & -\ 0.0324\ \log(POP) & +\ 0.0037\ \log(SEC) & -\ 0.0037\ \log(GDP60) \\ & & (0.0110) & (0.0019) & (0.0024) \end{array}$$

The Solow model assumes constant returns to scale, that is the coefficients of $\log(INV)$, $\log(POP)$, and $\log(SEC)$ should sum to zero. To choose δ, let us evaluate the gain of introducing the explanatory variables. Twice the estimated KLIC between our model and the one without regressors (but the intercept) divided by the sample size, is about 0.5. Fix δ^2 at 5% of this difference, that is 0.025. The LR test statistic has a value of $3\ 10^{-5}$, and the P-value is 0.45%.[3] Hence for any larger significance level the test concludes that the restriction is approximately valid. We also computed each of the three alternative test statistics H, S, and W. They agree with the LR statistic up to the tenth decimal. Hence all tests yield the same conclusion. In particular, the Wald model equivalence test asserts that the sum of the estimated coefficients of $\log(INV)$, $\log(POP)$, and $\log(SEC)$ is within 0.158 standard deviations of zero. Given a standard error of 0.011, this means that this sum is less than 0.0017 in absolute value.

Example 2: Testing a Consequence of Economic Theory. Anderson and Blundell (1983) estimates a flexible dynamic demand system on annual aggregate Canadian data by full information ML. They note that more restrictive models, as an autoregressive

[3]Matlab code to obtain the P-value of a model equivalence test is available from the author upon request.

model, a partial adjustment model, or a static model, are strongly rejected by significance tests. They also note that while homogeneity and symmetry restrictions are rejected within the static model, they are not within their dynamic setup. Their testing results are based on LR tests at 1% level and summarized in their Table 5. I focus on the results relative to their model "Dynamic: Price Index (4)." To choose the value of δ^2, I consider as a base model the static model.[4] The gain of modeling dynamics, as measured by twice the estimated KLIC between the dynamic and static models divided by the sample size, is 5.51. Consider testing for homogeneity, that is four restrictions, and choose δ^2 liberally as 4, that is about 73% of the gain of modeling dynamics. The test statistic is 10.2 and the corresponding P-value 72.4%. Hence one should allow the test to accept falsely homogeneity in almost three fourth of the cases to accept model equivalence. Therefore, the test does not confirm that homogeneity approximately holds. We note that the equivalence LR test would yield the same outcome at 10% level for any δ^2 smaller than 16.43, that is three times the gain of modeling dynamics. When simultaneously testing homogeneity and symmetry (ten restrictions), and assuming we also chose $\delta^2 = 4$, the test statistic equals 23.2 and the P-value is 92.3%. So, while significance tests at 1% level fail to reject either sets of restrictions, model equivalence tests fail short to accept that homogeneity and symmetry approximately hold.

Example 3 : Testing Exogeneity. Lillard and Aigner's (1984) analysis of time-of-day electricity demand rely on a two-equations triangular system in which the first equation explains air conditioning appliance ownership and the second explains electricity demand. The appliance ownership variables enter the second equation as explanatory variables and are exogenous if the first equation error ε is uncorrelated with each of the two components k and r of the second equation error. These correlations are denoted by $\rho_{k\varepsilon}$ and $\rho_{r\varepsilon}$ respectively. The system is estimated by full information ML. This application is also considered by Andrews (1989), which allows to compare his findings with ours. As

[4]One could also consider as an inadequate model the autoregressive or the partial adjustment model that are both judged inadequate by the authors. This would only strengthen our conclusions.

Andrews, I focus on the "Rate B all customers" results. Lillard and Aigner used a LR significance test to argue that the correlation coefficients are jointly insignificant at the 5% error level. Andrews argue that this conclusion does not seem warranted based on the estimated inverse power measures of the two univariate significance Wald tests. Based on the implied standard errors, choosing $\delta^2 = 1$ implies that model equivalence is declared here whenever $|\rho_{k\varepsilon}| \leq 0.207$ and $|\rho_{r\varepsilon}| \leq 0.246$ respectively, which are relatively large correlations.[5] The Wald statistics are 1.976 and 1.527 and the corresponding P-values for model equivalence tests are respectively 65% and 58.1%. Thus we conclude that each restriction is not approximately valid. If we consider the LR model equivalence test of the joint restrictions with $\delta^2 = 1$, the test statistic equals 1.8 and the P-value is 43.4%. Hence we conclude that approximate exogeneity does not hold.

4 Asymptotics

We now turn to the formal properties of our test. It is well known that in general there is no asymptotically uniformly most powerful tests in parametric models. It is then necessary to adopt a local approach in the search of optimal tests, see e.g. Lehmann and Romano (2005) for the local analysis of two-sided significance tests and equivalence tests of univariate restrictions. That is, for an arbitrary $\bar{\theta}$ such that $g(\bar{\theta}) = 0$, we analyse the asymptotic properties of our tests on the restricted set $\left\{\bar{\theta} + hn^{-1/2}, h \in \mathbb{R}^p, \|h\| \leq M\right\}$. We adopt two criteria for evaluating our model equivalence tests, local maximin optimality and local power in the class of tests invariant to orthogonal transformations. Local maximin optimality is used to characterize the classical trinity of significance tests in parametric models with multivariate parameters, see e.g. Borovkov (1998) and Lehmann and Romano (2005). The latters note that the maximin property may not be compelling for multiparameter significance hypotheses because the distant hypothesis can be defined through different norms. In the model equivalence framework however, the form of the

[5]Since Lillard and Aigner used the parametrization $\tan(\rho_{k\varepsilon}\pi/2)$ and $\tan(\rho_{r\varepsilon}\pi/2)$, stating results in terms of the correlations themselves requires a nonlinear transformation.

distant hypothesis is dictated by the considered hypotheses. Asymptotic invariance to linear trandformations is asuumed by Choi, Hall, and Schick (1998) to show local asymptotic optimality of two-sided significance tests of multivariate parameters. Here we also restrict to the class of tests invariant to orthogonal transformations of the parameter space, which is a mild requirement fulfilled even by the Wald test. We found that the model equivalence tests are locally asymptotically maximin, and as a consequence are locally asymptotically unbiased and most powerful against $g(\theta_0) = 0$. They also are locally asymptotically UMP invariant. In the case of univariate restrictions, the local asymptotic UMP property holds without invariance restriction.

Since model equivalence tests and significance tests are based on the same statistics, one may think that such results can be derived easily from existing ones. This is however not true. A first difficulty comes from the fact that avalaible results on significance tests assume that the parametrization of $\theta_0 = (\theta_{01}, \theta_{02})$ is such that the restrictions completely determine the value of θ_{01}, see e.g. Lehmann and Romano (2005) and Choi, Hall, and Schick (1998). While such a local (possibly nonlinear) reparametrization is always feasible, it is at odds with the invariance principle we want to invoke. Moreover, in the study of significance tests, the components of θ_{02} are treated as "nuisance parameters," because they are unconstrained under the null hypothesis. This leads authors to base their analysis on the "effective score," see e.g. Choi, Hall, and Schick (1998). By contrast, the model equivalence hypotheses involve the whole parameter vector, and there is strictly speaking no nuisance parameters. As a result, instead of considering the score test, we focus on the Hausman-Wald formulation in our theoretical analysis, which is extremely tractable because it directly involves the parameter vector.

To study the properties of our tests, we define $K_n^{LR}(\gamma) = \{\theta_0 : 2\ KLIC \leq \gamma^2/n\}$ for any $\gamma > 0$ and $\partial K_n^{LR}(\gamma) = \{\theta_0 : 2\ KLIC = \gamma^2/n\}$. For $J = H, S, W$, we define similarly $K_n^J(\gamma)$ and $\partial K_n^J(\gamma)$ as the similar sets based on the different formulations detailed above.

Theorem 4.1 *Suppose $X_1, \ldots, X_n$ are i.i.d. according to P_{θ_0}, $\theta_0 \in \Theta$, and that Assumptions A and B hold.*

(A) Let φ_n be a pointwise asymptotically level α tests sequence, that is

$$\limsup_{n\to\infty} \mathbb{E}_\theta \varphi_n \leq \alpha \quad \forall\ \theta \in H_n^J, \quad J = LR, H, S, \text{ or } W.$$

Let $\bar{\theta} \in \Theta$ be an arbitrary parameter such $g(\bar{\theta}) = 0$, $M > 0$ arbitrary large, and $\mathcal{N}(\bar{\theta}, M) = \left\{\bar{\theta} + hn^{-1/2},\ h \in \mathbb{R}^p,\ \|h\| \leq M\right\}$.

1. *For $\gamma^2 < \delta^2$, $J = LR, H, S,$ or W,*

$$\limsup_{n\to\infty} \inf_{\theta_0 \in K_n^J(\gamma) \cap \mathcal{N}(\bar{\theta}, M)} \mathbb{E}_{\theta_0} \varphi_n \leq \Pr\left[\chi_r^2(\gamma^2) \leq c_{\alpha,r,\delta^2}\right]. \tag{4.5}$$

2. *Assume φ_n is invariant to orthogonal transformations of the parameter space. Then for all $\gamma^2 < \delta^2$ and all $\theta_0 \in \partial K_n^J(\gamma) \cap \mathcal{N}(\bar{\theta}, M)$, $J = LR, H, S,$ or W,*

$$\limsup_{n\to\infty} \mathbb{E}_{\theta_0} \varphi_n \leq \Pr\left[\chi_r^2(\gamma^2) \leq c_{\alpha,r,\delta^2}\right]. \tag{4.6}$$

(B) Each tests sequence π_n^J, $J = LR, H, S,$ or W,

1. *is pointwise asymptotically level α,*

2. *is locally asymptotically maximin, in the sense that Inequality (4.5) is an equality. As a consequence, each tests sequence is locally asymptotically unbiased and locally most powerful against $\theta_0 = \bar{\theta}$.*

3. *is locally asymptotically UMP among tests invariant to orthogonal transformations, i.e. Inequality (4.6) is an equality.*

In our analysis, we rely on the local asymptotic normality of the likelihood ratio and the asymptotic equivalent experiments setting, see Le Cam and Lo Yang (2000) and Van der Vaart (1998). This reduces the problem to one of finding an optimal test in a normal experiment when we observe a sample of size one from $Z \sim N(\mu, \Sigma)$ and want to test

$$H\ : \mu'\Sigma^{-1/2} P \Sigma^{-1/2}\mu \geq \delta^2 \qquad \text{against} \qquad K\ : \mu'\Sigma^{-1/2} P \Sigma^{-1/2}\mu < \delta^2\ ,$$

where P is a known orthogonal projection matrix of rank r. Because this is of independent interest, we state here the result that characterizes the maximin test of H against

$$K(\gamma): \ \mu'\Sigma^{-1/2}P\Sigma^{-1/2}\mu \leq \gamma^2 < \delta^2$$

in the limiting normal experiment. The test $\pi(z)$ rejects H when $z'\Sigma^{-1/2}P\Sigma^{-1/2}z < c_{\alpha,r,\delta^2}$, where c_{α,r,δ^2} is the α quantile of a $\chi^2_r(\delta^2)$ distribution. Since the test is maximin, it is necessarily admissible and unbiased. Moreover, since it is independent of γ^2, it must be most powerful against $\mu = 0$. Finally, since it is also invariant to orthogonal transformations of the parameter space, it must be UMP invariant. These properties directly translate in equivalent local asymptotic properties for our tests.

Lemma 4.2 *Consider testing H against K from one observation z from $Z \in \mathbb{R}^p$ which is multivariate normal $N(\mu, \Sigma)$ with unknown mean μ and known nonsingular covariance matrix Σ. Then $\pi(z)$ is of level α, and among level α tests, π is maximin against $K(\gamma)$ with guaranteed power* $\Pr\left[\chi^2_r(\gamma^2) \leq c_{\alpha,r,\delta^2}\right]$.

We now consider the particular case of univariate restrictions, for which stronger results hold. Assume that $g(\cdot)$ is real-valued and can take positive and negative values. In that situation, our model equivalence tests are equivalent to the score and Wald procedures proposed by Romano (2004). His main test π^R_n rejects $|g(\theta)| \geq \tilde{\delta}/\sqrt{n}$ in favor of $|g(\theta)| < \tilde{\delta}/\sqrt{n}$ if $n^{1/2}|g(\widehat{\theta}_n)| \leq C(\alpha, \tilde{\delta}, \widehat{\sigma}_n)$, where $\widehat{\sigma}^2_n = \nabla'_\theta g(\widehat{\theta}_n) I^{-1}(\widehat{\theta}_n) \nabla_\theta g(\widehat{\theta}_n) = \sigma^2 + o_p(1)$ and $\sigma^2 = \nabla'_\theta g(\theta_0) I^{-1}(\theta_0) \nabla_\theta g(\theta_0)$. Here $C = C(\alpha, \delta, \sigma)$ is defined as the solution of

$$\Phi\left(\frac{C-\delta}{\sigma}\right) - \Phi\left(\frac{-C-\delta}{\sigma}\right) = \alpha\,,$$

$\Phi(\cdot)$ being the cumulative distribution function of a standard Gaussian real random variable. As

$$n^{1/2}|g(\widehat{\theta}_n)| \leq C(\alpha, \tilde{\delta}, \widehat{\sigma}_n) \Longleftrightarrow n\frac{g^2(\widehat{\theta}_n)}{\widehat{\sigma}^2_n} \leq \frac{C^2(\alpha, \tilde{\delta}, \widehat{\sigma}_n)}{\widehat{\sigma}^2_n} = C^2(\alpha, \frac{\tilde{\delta}}{\widehat{\sigma}_n}, 1)\,,$$

see Equation (6) in Romano (2005), and $C(\alpha, \delta, 1)$ is continuous in δ for any α,

$$C(\alpha, \frac{\tilde{\delta}}{\widehat{\sigma}_n}, 1) = C(\alpha, \frac{\tilde{\delta}}{\sigma}, 1) + o_p(1)\,.$$

Romano's test is then asymptotically equivalent to the one which rejects if

$$n\frac{g^2(\widehat{\theta}_n)}{\widehat{\sigma}_n^2} \leq C^2(\alpha, \frac{\tilde{\delta}}{\sigma}, 1)\,.$$

Now it is clear that $C^2(\alpha, \frac{\tilde{\delta}}{\sigma}, 1) = c_{\alpha,1,\tilde{\delta}^2/\sigma^2}$, so our Wald model equivalence test is asymptotically equivalent to π_n^R if we set $\delta^2 = \tilde{\delta}^2/\sigma^2$. Since our other model equivalence tests are also asymptotically equivalent the Wlad model equivalence test, the local asymptotic UMP property of Romano's test extends to each of our tests sequences. We here state this result without proof.

Corollary 4.3 *Assume that $g(\cdot)$ takes value in $\mathbb{R}$ and $g(\Theta)$ includes positive as well as negative values. Under the assumptions of Theorem 4.1, let φ_n be a pointwise asymptotically level α tests sequence, that is*

$$\limsup_{n\to\infty} \mathbb{E}_\theta \varphi_n \leq \alpha \quad \forall\ \theta \in H_n^J, \quad J = LR, H, S,\ \text{or } W.$$

Then for all $\gamma^2 < \delta^2$ and all $\theta_0 \in \partial K_n^J(\gamma) \cap \mathcal{N}(\bar{\theta}, M)$, $J = LR, H, S,$ or W,

$$\limsup_{n\to\infty} \mathbb{E}_{\theta_0} \varphi_n \leq \Pr\left[\chi_1^2(\gamma^2) \leq c_{\alpha,1,\delta^2}\right]\,. \tag{4.7}$$

Moreover, each tests sequence π_n^J, $J = LR, H, S,$ or W, is pointwise asymptotically level α and is locally asymptotically UMP, i.e. Inequality (4.7) is an equality.

5 Conclusion

We have proposed a theoretical framework to test whether some parameters restrictions are approximately valid in a parametric model. The framework is based on the Kullback-Leibler Information Criterion discrepancy, as is the standard LR significance test. The model equivalence hypothesis under test states that the discrepancy between the restricted and unrestricted model is smaller than some threshold that goes to zero as the sample size increases. We also investigated alternative formulation of this hypothesis. Of particular interest are the Hausman-Wald formulation, which evaluates the effect of the restrictions

on the whole parameter vector, and the Wald formulation, whose intuitive content yields equivalence intervals or regions that provide useful information on the parameters values that define the model equivalence hypothesis. Our likelihood-ratio model equivalence test, as well as its variants derived from alternative formulations of the hypothesis, have desirable optimality properties. Moreover we have shown through several examples that these tests are easy to apply and can prove useful in practical applications.

We focused on purpose on a parametric model which is well specified, i.e. contains the true data generating process, while the restrictions are not supposed to hold perfectly. This allowed us to obtain pretty strong theoretical results. Clearly one would like to extend the framework and model equivalence tests to contexts where the complete model could be misspecified. More crucially, extension to semiparametric models would be extremely useful in econometrics, and would allow in particular to propose equivalence tests for overidentification restrictions. The theoretical and practical properties of such tests will be explored in future research.

6 Proofs

Proof of Lemma 2.1. Assumption A implies that $\mathbb{E}_{\theta_0} l(X, \theta)$ is continuous in θ and attains its unique maximum at θ_0. Hence $2\ KLIC = O(n^{-1})$ implies that $\|\theta_0 - \theta_0^c\| = o(1)$. From a Taylor expansion and the information matrix equality,

$$\mathbb{E}_{\theta_0} l(X, \theta_0^c) \;=\; \mathbb{E}_{\theta_0} l(X, \theta_0) + (1/2)\,(\theta_0 - \theta_0^c)'\, I(\theta_0)\,(\theta_0 - \theta_0^c)\,(1 + o(1))\ .$$

Hence $\|\theta_0 - \theta_0^c\| = O(n^{-1/2})$ when $KLIC = O(n^{-1})$ since $I(\theta_0)$ is positive definite, and

$$2\ KLIC = (\theta_0 - \theta_0^c)'\, I(\theta_0)\,(\theta_0 - \theta_0^c)\,(1 + o(1))\ .$$

Similarly,

$$\begin{aligned} \mathbb{E}_{\theta_0} \nabla_\theta l(X, \theta_0^c) &= I(\theta_0)\,(\theta_0 - \theta_0^c)\,(1 + o(1)) \qquad (6.8) \\ \Rightarrow \mathbb{E}_{\theta_0}\left[\nabla'_\theta l(X, \theta_0^c)\right] I^{-1}(\theta_0) \mathbb{E}_{\theta_0}\left[\nabla_\theta l(X, \theta_0^c)\right] &= (\theta_0 - \theta_0^c)'\, I(\theta_0)\,(\theta_0 - \theta_0^c)\,(1 + o(1))\ , \end{aligned}$$

as $I^{-1}(\theta_0)$ exists by Assumption A(e). From Assumption B,

$$0 = g(\theta_0^c) = g(\theta_0) + \nabla_\theta' g(\theta_0)\,(\theta_0 - \theta_0^c)\,(1 + o(1))\ .$$

Let P_0 be the orthogonal projection matrix on $I^{-1/2}(\theta_0)\nabla_\theta g(\theta_0)$. Then

$$\begin{aligned} & g'(\theta_0)\left[\nabla_\theta' g(\theta_0) I^{-1}(\theta_0)\nabla_\theta g(\theta_0)\right]^{-1} g(\theta_0) \\ &= \ (\theta_0 - \theta_0^c)'\, I^{1/2}(\theta_0) P_0 I^{1/2}(\theta_0)\,(\theta_0 - \theta_0^c)\,(1 + o(1))\ . \end{aligned}$$

The constrained optimization problem for θ_0^c yields $\nabla_\theta \mathbb{E}_{\theta_0} l(X, \theta_0^c) = \nabla_\theta g(\theta_0^c)\lambda$ for some $\lambda \in \mathbb{R}^r$, so that $I^{-1/2}(\theta_0)\nabla_\theta g(\theta_0^c)\lambda = I^{1/2}(\theta_0)\,(\theta_0 - \theta_0^c)\,(1 + o(1))$ from (6.8). From Assumption B, $\nabla_\theta g(\theta_0^c) = \nabla_\theta g(\theta_0)\,(1 + o(1))$ and both matrices have the same rank. Combine these facts to obtain

$$(\theta_0 - \theta_0^c)'\, I^{1/2}(\theta_0) P_0 I^{1/2}(\theta_0)\,(\theta_0 - \theta_0^c) = (\theta_0 - \theta_0^c)'\, I(\theta_0)\,(\theta_0 - \theta_0^c)\,(1 + o(1))\ .$$

■

Proof of Lemma 4.2. Because Z can always be pre-multiplied by $\Sigma^{-1/2}$ to get an identity covariance matrix, there is no loss of generality to assume $\Sigma = \mathbf{I}_p$. Since P is an orthogonal projection matrix, there exists an orthogonal matrix A, i.e. $AA' = A'A = \mathbf{I}_p$, such that

$$A'PA = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad \text{for which} \quad X \equiv A'Z \sim N\left(A'\mu = \begin{pmatrix} \varepsilon_r \\ \varepsilon_l \end{pmatrix}, \mathbf{I}_p\right)\ .$$

Moreover, $\mu' P \mu = \mu' AA'PAA'\mu = \varepsilon_r'\varepsilon_r$, so that the hypotheses write

$$H\ : \varepsilon_r'\varepsilon_r \geq \delta^2 \qquad \text{against} \qquad K\ : \varepsilon_r'\varepsilon_r < \delta^2\ .$$

As X_r is sufficient for ε_r, we can restrict to tests based on it only. We aim to determine a minimax test of H against $K(\gamma)\ : \varepsilon_r'\varepsilon_r \leq \gamma^2$, which is a Bayes test under least favorable a priori distributions. Since the testing problem is invariant under orthogonal transformations, these distributions should also be invariant. Moreover, they should be concentrated on the boundary of the hypotheses. Therefore Q_δ, the uniform distribution

on the hypersphere $S(\delta)$ of radius δ, and Q_γ, defined similarly, are the least favorable a priori distributions. The most powerful Bayes test $\pi(x)$ of level α rejects H iff

$$\int_{S(\gamma)} \exp\left[-\frac{1}{2}(x_r-\varepsilon_r)'(x_r-\varepsilon_r)\right] dQ_\gamma(\varepsilon_r) > C \int_{S(\delta)} \exp\left[-\frac{1}{2}(x_r-\varepsilon_r)'(x_r-\varepsilon_r)\right] dQ_\delta(\varepsilon_r)$$

for some constant C. The left-hand side term writes

$$\exp\left[-\frac{1}{2}(x_r'x_r+\gamma^2)\right] \int_{S(\gamma)} \exp\left[x_r'\varepsilon_r\right] dQ_\gamma(\varepsilon_r) \ .$$

Denoting $e_x = x_r/\|x_r\|$, the above integral equals

$$\psi\left(\gamma\|x_r\|\right) = \int_{S(1)} \exp\left[\gamma\|x_r\|e_x\varepsilon_r\right] dQ_1(\varepsilon_r) = \int_{S(1)} \exp\left[\gamma\|x_r\|\varepsilon_{r1}\right] dQ_1(\varepsilon_r) \ ,$$

where ε_{r1} is the first component of ε_r. The function $\psi(\cdot)$ is strictly increasing with $\psi(0)=1$ and $\psi'(0)=0$. It is also logarithmically strictly convex. Indeed, for $t \neq u$ and $0<\lambda<1$

$$\begin{aligned}
\psi\left(\lambda t+(1-\lambda)u\right) &= \int_{S(1)} \left[\exp\left(t\varepsilon_{r1}\right)\right]^\lambda \left[\exp\left(u\varepsilon_{r1}\right)\right]^{1-\lambda} dQ_1(\varepsilon_r) \\
&< \left[\int_{S(1)} \exp\left(t\varepsilon_{r1}\right) dQ_1(\varepsilon_r)\right]^\lambda \left[\int_{S(1)} \exp\left(u\varepsilon_{r1}\right) dQ_1(\varepsilon_r)\right]^{1-\lambda} \\
&= \psi^\lambda(t)\,\psi^{1-\lambda}(u) \\
\Rightarrow \log\psi\left(\lambda t+(1-\lambda)u\right) &< \lambda\log\psi(t)+(1-\lambda)\log\psi(u) \ .
\end{aligned}$$

The rejection region of the test is

$$A\psi\left(\gamma\|x_r\|\right) > \psi\left(\delta\|x_r\|\right) \Leftrightarrow h\left(\|x_r\|\right) \equiv \log A + \log\psi\left(\gamma\|x_r\|\right) - \log\psi\left(\delta\|x_r\|\right) > 0 \ .$$

If $\log A \leq 0$, then $h(0)\leq 0$, and since $\log\psi(\cdot)$ is increasing, $h'(t)<0$ for all $t>0$, so that the above inequality holds for all $t>0$, which is clearly impossible if $\alpha<1$. Then it should be that $\log A>0$ and $h(0)>0$. Now there should be at least one $t_0>0$ such that $h(t_0)=0$ (since $\alpha<1$), and necessarily $h'(t_0)<0$. Since $\psi(\cdot)$ is logarithmically strictly convex and $\delta>\gamma$, $h''(t)<0$ for all $t>0$. Therefore $h'(t)<0$ for all $t>t_0$, that is t_0 is

unique. The test is then $\|x_r\|^2 < c$ for some constant c, which also writes

$$x' \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} x = z'Pz < c\,.$$

Hence the most powerful Bayes test of level α obtains for $c = c_{\alpha,r,\delta^2}$. Let us check that this test is minimax of level α. We have

$$\mathbb{E}_\mu \pi(Z) = \mathbb{P}\left[Z'PZ < c\right] = \mathbb{P}\left[\chi^2_r(\mu'P\mu) < c\right]\ .$$

As this probability is decreasing in $\mu'P\mu$ for each c,

$$\begin{aligned} \mathbb{E}_\mu \pi(Z) = \mathbb{P}\left[\chi^2_r(\mu'P\mu) < c\right] &\leq \mathbb{P}\left[\chi^2_r(\delta^2) < c\right] \qquad \text{for } \mu'P\mu \geq \delta^2 \\ \mathbb{E}_\mu \pi(Z) = \mathbb{P}\left[\chi^2_r(\mu'P\mu) < c\right] &\geq \mathbb{P}\left[\chi^2_r(\gamma^2) < c\right] \qquad \text{for } \mu'P\mu \leq \gamma^2\ , \end{aligned}$$

which yields

$$\sup_{\mu\in H} \mathbb{E}_\mu \pi(X) = \mathbb{E}_\mu \pi(X) \quad \forall \mu \in Q_\delta \quad \text{and} \qquad \inf_{\mu\in K(\gamma)} \mathbb{E}_\mu \pi(X) = \mathbb{E}_\mu \pi(X) \quad \forall \mu \in Q_\gamma\ .$$

Hence the test is minimax, see e.g. Borovkov (1998, Theorem 49.1), and is unbiased by definition of a minimax test. Since it is most powerful for testing H against $K(\gamma)$ under Q_δ and Q_γ and independent of γ, it is the most powerful test of H against $K(0)$. Moreover, it is also UMP among tests invariant to orthogonal transformations. ■

Proof of Theorem 4.1. We focus on the Hausman-Wald test, which is more convenient to deal with because it involves the basic parameter vector. We then explain briefly how the result extends to the other tests sequences.

Let $I_0 = I(\theta_0)$, P_0 be the orthogonal projection matrix on $I^{-1/2}(\theta_0)\nabla_\theta g(\theta_0)$, and define $\bar{I} = I(\bar{\theta})$ and $\bar{P}$ similarly.

i. Since $\mathbb{E}_{\theta_0} l(X,\theta_0) \geq \mathbb{E}_{\theta_0} l(X,\theta_0^c) \geq \mathbb{E}_{\theta_0} l(X,\bar{\theta})$,

$$0 \leq \mathbb{E}_{\theta_0} l(X,\theta_0) - \mathbb{E}_{\theta_0} l(X,\theta_0^c) \leq \mathbb{E}_{\theta_0} l(X,\theta_0) - \mathbb{E}_{\theta_0} l(X,\bar{\theta})\,. \tag{6.9}$$

Since $\|\theta_0 - \bar{\theta}\| = O(n^{-1/2})$, a Taylor expansion yields

$$\mathbb{E}_{\theta_0} l(X,\bar{\theta}) - \mathbb{E}_{\theta_0} l(X,\theta_0) = (1/2)\left(\theta_0 - \bar{\theta}\right)' I(\theta_0)\left(\theta_0 - \bar{\theta}\right)(1 + o(1)) = O(n^{-1})$$

uniformly in $\theta_0 \in \mathcal{N}(\bar{\theta}, M)$. Hence from (6.9) and another Taylor expansion

$$2\ KLIC = (\theta_0 - \theta_0^c)'\, I_0\, (\theta_0 - \theta_0^c)\, (1 + o(1)) = O(n^{-1}) \tag{6.10}$$

uniformly in θ_0. This shows that $\|\theta_0 - \theta_0^c\| = O(n^{-1/2})$. From Lemma 2.1's proof,

$$(\theta_0 - \theta_0^c)'\, I_0\, (\theta_0 - \theta_0^c) = (\theta_0 - \theta_0^c)'\, I_0^{1/2} P_0 I_0^{1/2}\, (\theta_0 - \theta_0^c)\, (1 + o(1))\ .$$

From the uniform continuity of $I(\theta)$ and $\nabla_\theta g(\theta)$ in $\mathcal{N}(\bar{\theta}, M)$,

$$(\theta_0 - \theta_0^c)'\, I_0\, (\theta_0 - \theta_0^c) = (\theta_0 - \theta_0^c)'\, \bar{I}^{1/2} \bar{P} \bar{I}^{1/2}\, (\theta_0 - \theta_0^c)\, (1 + o(1))\ . \tag{6.11}$$

By another Taylor expansion and the continuity of $\nabla'_\theta g(\cdot)$,

$$g(\theta_0^c) = 0 = g(\bar{\theta}) + \nabla'_\theta g(\bar{\theta})\left(\theta_0^c - \bar{\theta}\right) + o(\|\theta_0^c - \bar{\theta}\|) \Rightarrow \bar{P}\bar{I}^{1/2}\left(\theta_0^c - \bar{\theta}\right) = o(\|\theta_0^c - \bar{\theta}\|)\,.$$

Expand the right-hand side term of (6.11) to obtain

$$(\theta_0 - \theta_0^c)'\, I_0\, (\theta_0 - \theta_0^c)' = n^{-1} h' \bar{I} \bar{P} \bar{I} h\, (1 + o(1))\ , \tag{6.12}$$

uniformly in $\theta_0 \in \mathcal{N}(\bar{\theta}, M)$.

ii. Since the sequence of experiments $P^n_{\bar{\theta}+hn^{-1/2}}$ converges to a limiting normal experiment Z with unknown mean h and *known* covariance matrix $\bar{I}^{-1}$, it follows that we can approximate pointwise the power of any test φ_n by the power of a test in the limit experiment, see Van der Vaart (1998, Theorem 15.1) and Lehman and Romano (2005, Theorem 13.4.1). Since the limit hypothesis is $h'\bar{I}\bar{P}\bar{I}h < \delta^2$, apply Lemma 4.2 to deduce the bounds (4.5) and (4.6).

iii. Let $\Delta_n = n^{-1/2} \sum_{t=1}^n \nabla_\theta \log f(X_t; \bar{\theta})$. Under Assumptions A and B, standard results on maximum likelihood estimation, see e.g. Gourieroux and Monfort (1989), White (1994), Van der Vaart (1998), imply that under $\mathbb{P}^n_{\bar{\theta}}$

$$\sqrt{n}\left(\widehat{\theta}_n - \bar{\theta}\right) = -\bar{I}^{-1}\Delta_n + o_p(1)\,, \qquad \sqrt{n}\left(\widehat{\theta}^c_n - \bar{\theta}\right) = \bar{I}^{-1/2}\bar{M}\bar{I}^{1/2}\sqrt{n}\left(\widehat{\theta}_n - \bar{\theta}\right) + o_p(1)\,,$$

where $\bar{M} = \mathbf{I}_p - \bar{P}$. Under Assumption A, the model is differentiable in quadratic mean over Θ, see van der Vaart (1998, Lemma 7.6), and local asymptotic normality of the

log-likelihood ratio follows, that is

$$\sqrt{n}\ln\prod_{t=1}^{n}\frac{f_{\bar{\theta}+hn^{-1/2}}(X_t)}{f_{\bar{\theta}}(X_t)} = h'\Delta_n - h'\bar{I}h/2 + o_p(1) \qquad \forall h\in\mathbb{R}^p\,.$$

Since $\Delta_n \xrightarrow{d} N(0,\bar{I})$ under $\mathbb{P}^n_{\bar{\theta}}$, we obtain by Le Cam's third Lemma, see e.g. van der Vaart (1998), that for under $\mathbb{P}^n_{\bar{\theta}+hn^{-1/2}}$ and for any $h\in\mathbb{R}^p$

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}_n-\bar{\theta}\right) &\equiv \tau_n = Z+o_p(1)\,, \qquad Z\sim N(h,\bar{I}^{-1})\,,\\ \sqrt{n}\left(\widehat{\theta}^c_n-\bar{\theta}\right) &= \bar{I}^{-1/2}\bar{M}\bar{I}^{1/2}\tau_n+o_p(1)\,.\end{aligned}$$

This yields $\sqrt{n}\left(\widehat{\theta}_n-\widehat{\theta}^c_n\right)=\bar{I}^{-1/2}\bar{P}\bar{I}^{1/2}\tau_n$ for any $h\in\mathbb{R}^p$. Since $I(\widehat{\theta}_n)=\bar{I}+o_p(1)$, then for any $h\in\mathbb{R}^p$

$$n\left(\widehat{\theta}_n-\widehat{\theta}^c_n\right)'I_n(\widehat{\theta}_n)\left(\widehat{\theta}_n-\widehat{\theta}^c_n\right)=n\left(\widehat{\theta}_n-\widehat{\theta}^c_n\right)'\bar{I}\left(\widehat{\theta}_n-\widehat{\theta}^c_n\right)+o_p(1)=\tau_n\bar{I}^{1/2}\bar{P}\bar{I}^{1/2}\tau_n+o_p(1)\,.$$

iv. Consider $\pi(\tau_n)$, where π is the test defined in Lemma 4.2. Then $\mathbb{E}_{\bar{\theta}+hn^{-1/2}}\pi^H_n = \mathbb{E}_{\bar{\theta}+hn^{-1/2}}\pi(\tau_n)+o(1)$ pointwise in $h\in\mathbb{R}^p$ and $\tau_n\bar{I}^{1/2}\bar{P}\bar{I}^{1/2}\tau_n$ is for any $h\in\mathbb{R}^p$ asymptotically equivalent to a $\chi^2_r(h'\bar{I}^{1/2}\bar{P}\bar{I}^{1/2}h)$, see Rao and Mitra (1971, Lemma 9.12). As $\pi(\tau_n)$ test rejects H^H_n when $\tau_n\bar{I}^{1/2}\bar{P}\bar{I}^{1/2}\tau_n < c_{\alpha,r,\delta^2}$,

$$\mathbb{E}_{\bar{\theta}+hn^{-1/2}}\pi(\tau_n)=\mathbb{P}\left[\tau_n\bar{I}^{1/2}\bar{P}\bar{I}^{1/2}\tau_n<c_{\alpha,r,\delta^2}\right]\to\mathbb{P}\left[\chi^2_r(h'\bar{I}^{1/2}\bar{P}\bar{I}^{1/2}h)<c_{\alpha,r,\delta^2}\right]\,.$$

In particular, $\pi(\tau_n)$ and thus π^H_n are locally pointwise asymptotic level α.

Since π is Bayesian of level α for a priori measures Q_δ and Q_γ and

$$\mathbb{E}_{Q_\gamma}\pi(\tau_n)=\int_{S(\gamma)}\mathbb{E}_{\bar{\theta}+hn^{-1/2}}\pi(\tau_n)\,dQ_\gamma\to\mathbb{E}_{Q_\gamma}\pi(Z)$$

by the Lebesgue dominated convergence theorem, $\pi(\tau_n)$ and thus π^H_n are also asymptotically Bayesian level α for the same a priori measures.

For any other test sequence φ_n of asymptotically Bayesian level α,

$$\limsup_{n\to\infty}\inf_{K(\gamma)}\mathbb{E}_{\bar{\theta}+hn^{-1/2}}\varphi_n\le\limsup_{n\to\infty}\mathbb{E}_{Q_\gamma}\varphi_n\le\limsup_{n\to\infty}\mathbb{E}_{Q_\gamma}\pi(\tau_n)\,.$$

But $\limsup_{n\to\infty} \mathbb{E}_{Q_\gamma}\pi(\tau_n) = \mathbb{E}_{Q_\gamma}\pi(Z) = \inf_{K(\gamma)} \mathbb{E}_h\pi(Z) = \lim_{n\to\infty} \inf_{K(\gamma)} \mathbb{E}_{\bar\theta+hn^{-1/2}}\pi(\tau_n)$. Gathering results,

$$\liminf_{n\to\infty} \left(\inf_{K(\gamma)} \mathbb{E}_{\bar\theta+hn^{-1/2}}\pi(\tau_n) - \inf_{K(\gamma)} \mathbb{E}_{\bar\theta+hn^{-1/2}}\varphi_n \right) \geq 0\,,$$

which shows that $\pi(\tau_n)$ and thus π_n^H are locally asymptotically maximin.

Consider a test sequence φ_n of pointwise asymptotic level α and invariant to orthogonal transformations. Then for any γ and any $h \in S(\gamma)$

$$\limsup_{n\to\infty} \mathbb{E}_{\bar\theta+hn^{-1/2}}\varphi_n \leq \limsup_{n\to\infty} \mathbb{E}_{Q_\gamma}\varphi_n \leq \limsup_{n\to\infty} \mathbb{E}_{Q_\gamma}\pi(\tau_n) = \lim_{n\to\infty} \mathbb{E}_{\bar\theta+hn^{-1/2}}\pi(\tau_n)\,,$$

so that $\pi(\tau_n)$ and thus π_n^H are locally asymptotically UMP among invariant tests.

From (6.12) and the continuity of the power function of $\pi(\tau_n)$, deduce that the same local asymptotic properties hold for $\pi(\tau_n)$, and thus π_n^H, as tests of H_n^H against $K_n^H(\gamma)$. Finally note that for θ_0 such that $n^{1/2}\min_{g(\theta)=0}\|\theta_0-\theta\| \to \infty$, $n^{1/2}\tau_n \to \infty$ and the power of both tests tends pointwise to zero.

v. To extend the result to the LR test, use (6.10) to deduce that limits of infima on $K_n^H \cap \mathcal{N}(\bar\theta, M)$ are equal to limits of infima $K_n^{LR} \cap \mathcal{N}(\bar\theta, M)$. The local asymptotic equivalence of the LR test follows easily from the local asymptotic equivalence of the LR statistic to H, which follows by standard arguments, see e.g. Van der Vaart (1998). The result extends to the other tests following the same lines. ■

REFERENCES

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *Proceedings of the Second International Symposium on Information Theory*, B.N. Petrov and F. Csaki eds., 267–281. Akademiai Kiodo: Budapest.

Anderson, G., and Blundell, R. (1983). Testing Restrictions in a Flexible Dynamic Demand System: An Application to Consumers' Expenditure in Canada. *Rev. Econ. Stud.* 50(3), 397–410.

Andrews, D.W.K. (1989). Power in Econometric Applications. *Econometrica* 57, 1059–1090.

Andrews, D.W.K. (1994). The Large Sample Correspondence between Classical Hypotheses Tests and Bayesian Posterior Odds Tests. *Econometrica* 62(5), 1207–1232.

Arrow, K. (1960). Decision Theory and the Choice of a Level of Significance for the T-test. *Contributions to Probability and Statistics*, Olkin and al. eds., 70–78. Stanford Unievrsity Press: Stanford.

Berger, J.O., and Sellker, T. (1987). Testing a Point Null Hypothesis: The irreconciliability of P Values and Evidence. *J. Amer. Statist. Assoc.* 82(397), 112–122.

Berkson, J. (1942). Tests of Significance Considered as Evidence. *J. Amer. Statist. Assoc.* 37(219), 325–335.

Borovkov, A.A. (1998). *Mathematical Statistics*. Overseas Publishers Association: Amsterdam.

Choi, S., Hall, W.J., and Schick, A. (1998). Asymptotically Uniformly Most Powerful Tests in Parametric and Semiparametric Models. *Ann. Statist.* 24(2), 841–861.

Cox, D.R. (1958). *Planning of Experiments*. Wiley & Sons: New-York.

Dette, H., and Munk, A. (1998). Validation of Linear Regression Models. *Ann. Statist.* 26(2), 778–800.

Good, I.J. (1981). Some Logic and History of Hypothesis Testing. *Philosophical Foundations of Economics*, J.C. Pitt ed., 149–174.

Good, I.J. (1983). *Good Thinking: The Foundations of Probability and Its Applications*. University of Minessota Press: Minneapolis.

Good, I.J. (1992). The Bayes/Non-Bayes Compromise: A Brief Review. *J. Amer. Statist. Assoc.* 87(419), 597–606.

Gourieroux, C., and Monfort. A. (1989). *Statistics and Econometric Models*. Cambridge University Press: Cambridge.

Gregory, A.W., and Veall, M.R. (1985). Formulating Wald Tests of Nonlinear Restrictions. *Econometrica* 53, 1465–1468.

Hodges, J.L., and Lehmann, E.L. (1954). Testing the Approximate Validity of Statistical Hypotheses. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 16(2), 261–268.

Hoenig, J.M. and Heisey, D.M. (2001). The Abuse of Power: the Pervasive Fallacy of Power Calculations for Data Analysis. *Amer. Statist.* 55, 19–24.

Jeffreys, H. (1939). *Theory of Probability*. Clarendon Press: Oxford, U.K.

Kass, R.E., and Raftery, A.E. (1995). Bayes Factors. *J. Amer. Statist. Assoc.* 90(430), 773–795.

King, R.G., and R. Levine (1993). Finance and Growth: Schumpeter Might Be Right. *Quart. J. Econ.* **108 (3)**, 717–737.

Le Cam, L.M., and Lo Yang, G. (2000). *Asymptotics in Statistics*. Springer: New York.

Leamer, E.E. (1988). Things That Bother Me. *Econ. Rec.* 64, 331–335.

Lehmann, E. L. (1958). Significance Level and Power. *Annals of Math. Statist.* 29 (4), 1167–1176

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

Lillard, L.A., and Aigner, D.J. (1984). Time-of-Day Electricty Consumption Response to Temperature and the ownership of Air Conditioning Appliances. *J. Bus. Econ. Statist.* 2(1), 40–53.

Liu, Z., and T. Stengos (1999). Non-Linearities in Cross-Country Growth Regressions: a Semiparametric Approach. *J. Appl. Econometrics* **14 (5)**, 527–538.

Mankiw, N.G., D. Romer, and D.N. Weil (1992). A Contribution to the Empirics of Economic Growth. *Quart. J. Econ.* **107 (2)**, 407–437.

McCloskey, D.N. (1985). The Loss Function Has Been Mislaid: the Rhetoric of Significance Tests. *Amer. Econ. Rev.* 75, 201–205.

Rao, C.R., and Mitra, S.K. (1971). *Generalized Inverse of Matrices and its Applications*. Wiley & Sons: New York.

Romano, J.P. (2005). Optimal Testing of Equivalence Hypotheses. *Ann. Statist.* 33, 1036–1047.

Rosenblatt, J. (1962). Testing Approximate Hypotheses in the Composite Case. *Ann. Math. Statist.* 33, 1356–1364.

Sawa, T. (1978). Information Criteria for Discriminating Among Alternatives Regression Models. *Econometrica* 46(6), 1273–91.

Sen, S. (2001). Statistical Issues in Bioequivalence. *Statist. Med.* 20, 2785–2799.

Sin, C.Y., and White, H. (1996). Information Criteria for Selecting Possibly Misspecified Parametric Models. *J. Econometrics* 71, 207–225.

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

Vuong, Q.H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica* **57(2)**, 301-333.

Wellek, S. (2003). *Testing Statistical Hypotheses of Equivalence*. Chapman and Hall/CRC: New York.

White, H. (1994). *Estimation, Inference, and Specification Analysis*. Cambridge University Press: New York.