N° TSE-379

February 2013

Model Equivalence Tests in a Parametric Framework

Pascal Lavergne

Model Equivalence Tests in a Parametric Framework

Pascal Lavergne, Toulouse School of Economics

This version: February 2013

Abstract

In empirical research, one commonly aims to obtain evidence in favor 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 based on the Kullback-Leibler information 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.

I acknowledge financial support from SSHRC. Thanks to my colleagues Jean-Pierre Florens, Thierry Magnac, and to three referees, for their useful comments. 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 obtain evidence in favor of restrictions on parameters. 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, see e.g. Godfrey (1988). While in a test of significance, the researcher is typically hoping that the null hypothesis of insignificance will be rejected, in 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. While it has been early acknowledged that applied researchers are often looking for positive evidence *in favor* of restrictions, there seems to be no clear consensus on how to provide such evidence. Berkson (1942) argues that the p-value of a significance test can be used as evidential measure in favor of the null hypothesis. Some authors instead favor Bayes factors as introduced by Jeffreys (1961), see Kass and Raftery (1995) and the references therein. Good (1983, 1992) advocates for a compromise of Bayesian and non Bayesian approaches. Casella and Berger (1987) show that it is possible to reconcile p-values and bayesian posterior probability in one-sided testing problems, but Berger and Sellke (1987) argue that this is difficult for two-sided tests. However, Andrews (1994) shows that under certain asymptotics there exists a correspondence between p-values and Bayesian posterior odds.

The goal of this work is to develop "classical" tests 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. Here the approximate validity of the restrictions of interest is considered as the *alternative* hypothesis to reflect where the burden of proof is. This is known in biostatistics as *equivalence testing*, see Lehman and Romano (2005), Wellek's monograph (2003), and Senn's review (2001). Another application of this principle is provided by Dette and Munk (1998) for specification testing. Finally, the approximate alternative hypothesis concentrates around the sharp restrictions of interest to formalize that we are interested in showing that our restrictions are close to be fulfilled, see Rosenblatt (1962) for an early example. In this vein, Romano (2005) considers assessing an *univariate* restriction on parameters of the form $\theta = 0$ through a setting where the alternative hypothesis of interest is a neighborhood of the hypothesis of interest that becomes narrower as sample size, and thus information, increases, see also Borovkov (1998) for related results. The related test yields a decision on 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, the framework developed here does not directly focus on the restrictions themselves, but on the consequences of imposing these restrictions. Following Akaike (1973) and Vuong (1989), I evaluate the effect of the restrictions as measured by the Kullback Leibler Information Criterion ($KLIC$), which is a natural divergence measure between the unrestricted model and the restricted one. Hence the alternative hypothesis of interest states that the $KLIC$ is less than a small tolerance. This allows to consider univariate as well as multivariate restrictions on parameters, which has not been dealt with in previous work. I derive a test based on the usual likelihood-ratio (LR) statistic, but that uses a decision rule different from the one 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. The procedure has desirable invariance properties. I label this approach *model equivalence testing*.

One may wonder whether and why a new approach, that considers an alternative approximate hypothesis, is needed. Many authors in statistics and economics have emphasized that considering an approximate hypothesis makes more sense and can be more interesting than a point hypothesis. The issue actually dates back at least to Berkson (1938, 1942). Cox (1958) writes that for a point null hypothesis "Exact truth of a null hypothesis is very unlikely except in a genuine uniformity trial." Berger and Delampady (1987) discuss "precise" hypotheses, which "in reality are better represented as tests of, say,

$$H_0 : \ |\theta - \theta_0| \leq \varepsilon \qquad \text{versus} \quad H_1 : \ |\theta - \theta_0| > \varepsilon,$$

where ε is "small"." In many practical cases, it seems that an approximate hypothesis is exactly what we want to consider. For instance, in demand analysis, do we expect observations to strictly conform to consumer theory or to be "close" to what is predicted by our theory? Considering as our alternative hypothesis one that states that the restrictions are "almost" valid allows to "flip" the usual null and alternative hypothesis. While this is not in line with common econometric practice, it is however in line with the well-known statistical principle in classical hypothesis testing that states that we should consider as the alternative hypothesis what we expect or would like to show. Hence, when one expects a parameter θ to be positive (e.g. an income demand elasticity), sound statistical practice considers the alternative hypothesis $H_a : \theta > 0$. Similarly, when one would like to show that a variable is pertinent, one considers that its coefficient $\theta \neq 0$ as the alternative hypothesis and entertains a significance test. But the contradiction appears when one wants to show that the variable is not pertinent, since one still keeps $\theta \neq 0$ as the alternative hypothesis. In that case, common practice not only forgets about the principle, but turns it upside down. Our choice of the alternative hypothesis thus acknowledges where the burden of proof is.

Could one use instead well-known procedures to address the same issue? Significance testing is well suited for rejecting a point null hypothesis. But there are many instances where the aim is indeed to show that the restrictions are (close to be) fulfilled, see our examples above and our illustrations below. A significance test entertained at usual nominal levels can never accept the null hypothesis, and thus cannot assess the validity of restrictions, even in an approximate sense. This is because it tunes the probability of type-I error, that is the odds of falsely rejecting the restrictions, but does not control the probability of type-II error, that is the probability of *falsely not rejecting them*. 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. Could confidence intervals or regions be used instead? As these are defined as sets of parameters values that cannot be rejected by a significance test, they do not provide a suitable answer either. Another advocated approach is to rely on power evaluation of significance tests. In particular, Andrews (1989) proposes approximations of the asymptotic inverse power function as an aid to interpret non significant outcomes. These 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). It is also known that asymptotic power approximations of the Wald test can be misleading in a nonlinear model, see Nelson and Savin (1990). Apart from these technicalities, *evaluating* the asymptotic power of a significance test of given level does not directly provide evidence in favor of the restrictions under consideration. Other issues surround post-experiment power calculations, as summarized by Hoenig and Heisey (2001).[1] To sum up, model equivalence testing is not a substitute or competitor of significance testing, but instead delivers inference in situations where the latter approach is not well suited.

I also investigate whether alternatives procedures to the LR model equivalence test can be considered. I show indeed that one can derive asymptotically equivalent formulations of the hypotheses, and I focus on three of these. 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 Wald approach and is similar to Romano's equivalence test in the case of an univariate restriction. The third one relies on a score approach, but is valid only under a more restrictive as-

[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."

sumption. Each formulation yields an alternative model equivalence test, and each test is asymptotically equivalent to the LR model equivalence test, though they do not share all its invariance properties. I show that the four related model equivalence tests are, within their respective testing framework, locally asymptotically maximin and locally asymptotically most powerful in the class of tests invariant to orthogonal transformations of the parameter. These optimality properties are the ones used to characterize the classical trinity of significance tests, see Borovkov (1998) or Lehmann and Romano (2005). When considering an univariate restriction, the proposed equivalence tests are asymptotically equivalent to the test proposed by Romano (2005), and thus are locally asymptotically uniformly most powerful. However, the general theoretical analysis sharply differs from existing work for reasons to be explained in Section 4.

The paper is organized as follows. In Section 2, I setup the main testing framework based on the $KLIC$, I derive the model equivalence LR test, and discuss implementation through examples. In Section 3, I propose asymptotically equivalent frameworks and tests and illustrate their use. In Section 4, I study the local asymptotic properties of the tests. In Section 5, I conclude by suggesting directions for future research.

2 $KLIC$-Based Testing

2.1 Framework and Test

Let us introduce the basic setup considered throughout this paper. To focus on the main issues, I deal with unconditional models, but the 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 $\{X_t, t = 1, \ldots n\}$ from X, whose probability density $f(\cdot, \theta_0)$ belongs to a parametric family of densities $\{f(\cdot, \theta) : \theta \in \Theta\}$. Denote by $\mathbb{E}_{\theta_0}$ the expectation when θ_0 is the parameter value. We are interested in assessing 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 $g(\cdot)$ and θ_0 only. Following Akaike (1973, 1974), Sawa (1978), and Vuong (1989), among others, consider as a measure of closeness 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 divergence measure is always positive and zero if and only if the restrictions perfectly hold, see Vuong (1989). Consider then as the *alternative hypothesis* to assess

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

I label it the *model equivalence hypothesis*. It does not entail that the restrictions perfectly hold, but that these restrictions are close to be valid. We will return to the interpretation of the model equivalence hypothesis several times later on. For now, let us note that the smaller the tolerance δ^2/n, the more stringent the hypothesis. If one accepts model equivalence for a particular tolerance, then the decision will be the same for any larger one. The *null hypothesis* is the complement of the alternative, that is

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

The vanishing tolerance acknowledges that the tolerance is small in a substantive sense. In practice, a small but fixed tolerance Δ^2 is typically chosen, so that one can set $\delta^2 = n\Delta^2$. This is how I will apply the test in subsequent illustrations. But because the fixed tolerance is so small, the asymptotics under a drifting tolerance δ^2/n will approximate the finite sample distribution of the test statistic better than the asymptotics under a fixed tolerance Δ^2. Considering a shrinking hypothesis is thus purely a theoretical but very useful device.[2] From a theory viewpoint, our shrinking hypothesis setup with tolerance δ^2/n puts us in the most difficult but manageable situation. Indeed, would the tolerance go towards zero faster than $n^{-1/2}$, all distributions in K_n^{LR} would be contiguous to some distributions in

[2] A similar approach is adopted in power analysis, where exploring a significance test's power under local alternatives gives a better picture of actual power for a small or moderate sample size.

H_n^{LR} and then would not be distinguishable from H_n^{LR}, see e.g. Lehmann and Romano (2005).[3]

The model equivalence test is based on the likelihood ratio (LR). 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 LR test statistic is $2\ LR_n = 2\left[L_n(\widehat{\theta}_n) - L_n(\widehat{\theta}_n^c)\right]$. The LR model equivalence test of H_n^{LR} against K_n^{LR} is $\pi_n^{LR} = \mathbb{I}\left[2\ LR_n < c_{\alpha,r,\delta^2}\right]$, where c_{α,r,δ^2} 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. While critical values are non-standard, they can be readily obtained from most statistical softwares, and are reported in Tables 1 to 6 for the test at 10% and 5%, δ^2 varying from 0.1 to 5, and r =1 to 6.

2.2 Choice of the Tolerance in Applications

The choice of the tolerance is key in our procedure. In practice, it is often easier to choose a fixed tolerance Δ^2 for the divergence we are ready to tolerate between the two models and embed this into the sequence of hypotheses K_n^{LR} by setting $\delta^2 = n\Delta^2$. In what follows, I illustrate through examples how the tolerance Δ^2 can be chosen in practice and the test implemented.

2.2.1 Linear Regression

Consider first a linear regression model

$$Y = X'\beta_0 + \varepsilon, \qquad \varepsilon | X \sim N(0,\sigma_0^2)\,,$$

[3] One should 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 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.

and restrictions $g(\beta_0) = 0$. Using the properties of the conditional expectation,

$$\begin{aligned}
\sigma_{0c}^2 &\equiv \mathbb{E}_{\theta_0} \left[Y - X'\beta_0^c\right]^2 = \mathbb{E}_{\theta_0} \left[Y - X'\beta_0\right]^2 + \mathbb{E}_{\theta_0} \left[X'\beta_0^c - X'\beta_0\right]^2 \\
&= \sigma_0^2 + (\beta_0^c - \beta_0)' \, \mathbb{E}(XX') \, (\beta_0^c - \beta_0) \; .
\end{aligned}$$

Since in that setup

$$2 \; KLIC = 2 \; \mathbb{E}_{\theta_0} \left[\frac{\log f(Y|X, \theta_0)}{\log f(Y|X, \theta_0^c)} \right] = \log \frac{\sigma_{0c}^2}{\sigma_0^2} \, ,$$

then for values of $\sigma_{0c}^2 - \sigma_0^2$ close to 0,

$$\begin{aligned}
2 \; KLIC &= \log \left(1 + \frac{\sigma_{0c}^2 - \sigma_0^2}{\sigma_0^2} \right) \approx \frac{\sigma_{0c}^2 - \sigma_0^2}{\sigma_0^2} \\
&= \frac{(\beta_0^c - \beta_0)' \, \mathbb{E}(XX') \, (\beta_0^c - \beta_0)}{\sigma_0^2} \; .
\end{aligned}$$

Hence 2 $KLIC$ measures the loss in explanatory power coming from imposing the constraint relative to the error's variance.

EXAMPLE 1: RESTRICTIONS FROM AN ECONOMIC HYPOTHESIS. *I consider here a cross-country regression in the spirit of Mankiw and al. (1992), using pooled data on 86 countries averaged over the 1960's, 1970's and 1980's 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; and two dummy variables D70 and D80, acting as fixed effects for the seventies and the eighties. OLS estimation yields*

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

The Solow model assumes constant returns to scale, that is the coefficients of $\log(INV)$, $\log(POP)$, *and* $\log(SEC)$ *should sum to zero. For our application, let* $\Delta^2 = 0.1$, *i.e.* $\delta^2 = 0.1 \times n$, *that is model equivalence is declared if the explanatory power lost by imposing this constraint is at most* 10% *of the error's variance. The LR test statistic has* *a value of* 3 10^{-5}*, and the p-value of the model equivalence test is* 10^{-8}*. Hence for any larger significance level the test concludes that the restriction is approximately valid.*[4]

Another way of running the test is "in the spirit of the p-value approach," but instead of letting the test's level varies, we let δ^2 vary for a given level and we formally define

$$\delta^2_{\text{inf}}(\alpha) = \inf \left\{ \delta^2 > 0 : 2\ LR_n < c_{\alpha,r,\delta^2} \right\} .$$

This provide a useful benchmark, since practitioners may agree in some instances on whether $\Delta^2_{\text{inf}}(\alpha) = \delta^2_{\text{inf}}(\alpha)/n$ is close enough to zero or substantially different. Example 1 provides an illustration.

EXAMPLE 1 (CONTINUED): *In our application,* $\delta^2_{\text{inf}}(1\%) = 0$*, because the test statistic is smaller than* $c_{0.01,1,0}$*, that is the LR model equivalence test accepts model equivalence at a 1% level for any* $\delta^2 > 0$*. This gives strong evidence in favor of the approximate validity of the constant returns to scale hypothesis.*

2.2.2 The General Case

While the precise interpretation of 2 $KLIC$ may be case dependent, as exemplified above, some general remarks can be made. The divergence is a unitless quantity since it depends only on the ratio $f(\cdot,\theta_0)/f(\cdot,\theta_0^c)$. If this ratio is close to one uniformly in x, as it should be if the two models are close, then

$$\begin{aligned} 2\ KLIC &= -2\ \mathbb{E}_{\theta_0}\left[\log\left(1+\frac{f(X,\theta_0^c)-f(X,\theta_0)}{f(X,\theta_0)}\right)\right] \\ &= -2\left\{\mathbb{E}_{\theta_0}\left[\frac{f(X,\theta_0^c)-f(X,\theta_0)}{f(X,\theta_0)}\right]-\frac{1}{2}\mathbb{E}_{\theta_0}\left[\left(\frac{f(X,\theta_0^c)-f(X,\theta_0)}{f(X,\theta_0)}\right)^2\right](1+o(1))\right\} \\ &= \mathbb{E}_{\theta_0}\left[\left(\frac{f(X,\theta_0^c)-f(X,\theta_0)}{f(X,\theta_0)}\right)^2\right](1+o(1))\,. \end{aligned} \tag{2.2}$$

The divergence measures the expected squared proportional difference between distributions and is thus a squared percentage. Hence the tolerance Δ can be seen as a percentage.[5]

[4] R and Matlab codes to obtain the p-value of a model equivalence test are available from the author upon request.

[5] The previous interpretation for the linear regression model as the loss of fit in proportion of the error's variance is easily reconciled with the current one, by noting that the variance is expressed in squared

Example 2: Restrictions from a Consequence of Economic Theory. *Anderson and Blundell (1983) estimated a flexible dynamic demand system by full information ML on first-differenced budget shares using annual aggregate Canadian data. They note that more restrictive models, namely an autoregressive model, a partial adjustment model, and a static model, are strongly rejected by significance tests, see their Table 3. They also note that while homogeneity and symmetry restrictions are rejected for these restrictive models, they are not within their general 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 labeled "Dynamic: Price Index (10)."*

Consider homogeneity, that is four restrictions, and let us assume that one fix a tolerance $\Delta^2 = (50\%)^2$ *for the chosen divergence measure. The test statistic is 10.6 and the corresponding p-value is 47.01%. Therefore, the test does not confirm that homogeneity approximately holds. When considering simultaneously homogeneity and symmetry (ten restrictions), and assuming we chose the same tolerance, the test statistic equals 24.6 and the p-value is 82.78%.*

To determine whether a particular percentage, such as about 50% in the previous example, is sensible, it may be helpful to have in mind an upper bound for the divergence. Theoretical upper bounds can be derived, see for instance Borovkov (1998, Section 31, Theorem 3A), but might be not very useful in practice. In any application however, it is often easy to determine such a bound by considering the $KLIC$ between the complete model and a model that has already been judged inadequate on economic or statistical grounds. For instance, if a model has been strongly rejected by a significance test, then one can confidently assess that the divergence between this model and the complete one is large.[6]

Example 2 (continued): *The gain of the general dynamic structure compared to the autoregressive model is* $1.5063 = (122.73\%)^2$*, as estimated from the data.*[7] *Now for ho-*

units.

[6] This illustrates that model equivalence testing is by no means a substitute of significance testing, on the contrary the two approaches deliver useful complementary information.

[7] This is twice the difference of the log-likelihood of the maintained model, 686.2, and the one of the autoregressive model, 662.1, divided by the sample size, 32.

mogeneity and symmetry (ten restrictions), $\delta^2_{\inf}(1\%) = 42.13$ *and* $\Delta^2_{\inf}(1\%) = (114.74\%)^2$*. Hence, to confirm that these restrictions approximately hold, we would need to forsake almost the whole gain of modeling general dynamics. So, while significance tests fail to reject these restrictions, model equivalence tests fail short to accept that homogeneity and symmetry approximately hold.*

To sum up, three pieces of information can guide us in the practical choice of the tolerance: (i) the interpretation of $2KLIC$ as the fit lost by imposing the constraint (ii) the divergence between our general model and a model that is known to be inadequate, (iii) the determination of the value of the tolerance for which the the model equivalence test's outcome changes. We will see below that alternative expressions of the model equivalence hypothesis can shed more light on the interpretation and practical choice of the tolerance.

To complete our understanding, it is useful to know how the power of the test varies with the tolerance. 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. The power is increasing in, and pretty sensitive to, δ^2 and α. It is seen that the power is always maximum when $KLIC = 0$, that is when the restrictions perfectly hold, but never attains one. 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. As will be shown, no test can achieve a larger local asymptotic power at zero. Since the test statistic is the same as in a significance test, we could 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 $\Lambda(\gamma^2) = \Pr\left[\pi_n^{LR} | 2\, n\, KLIC = \gamma^2\right]$ is the (normalized) 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 $2\, KLIC = \delta^2/n$. The model equivalence approach however relies on a precise characterization of the approximate hypothesis and is much more direct.

3 Alternative Tests

3.1 Testing Frameworks

While $KLIC$ is a classical divergence measure between models, it is just one among several. We have noted that the chi-squared distance (2.2) is equivalent to $KLIC$ for two "close" models. Let us formalize such a concept of equivalence.

Definition 1 *1. $d(g,\theta_0)$ is a divergence measure between $f(X,\theta_0)$ and the model constrained by $g(\theta)=0$, as given by $f(X,\theta_0^c)$, if $d(g,\theta_0)\geq 0$ with equality iff $g(\theta_0)=0$.*
2. We say that two divergence measures d_i, $i=1,2$, are locally equivalent if, under any drifting sequence of parameters θ_{0n} , $n\geq 1$, such that $d_i(g,\theta_{0n})=o(1)$ or $d_j(g,\theta_{0n})=o(1)$, we have $d_i(g,\theta_{0n})=d_j(g,\theta_{0n})(1+o(1))$.

Most usual divergences, such as the chi-squared distance (2.2) or Hellinger's distance (up to a multiplicative factor 4), are locally equivalent to $KLIC$, see e.g. Borovkov (1998). This entails first that there is no "best" divergence to construct a testing framework, and second that we may consider other locally equivalent divergences. In what follows, I focus on alternative divergences that yield familiar testing frameworks. Denote by ∇_θ differentiation with respect to θ and by $\nabla_{\theta,\theta'}$ second differentiation, and let us make the following standard 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 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 $\|\nabla_{\theta,\theta'}l(x,\theta)\|<\bar{l}(x)$ and $\mathbb{E}_\theta\bar{l}^2(X)<\infty$ uniformly over a neighborhood of θ_0. (e) The information matrix $I(\theta)\equiv\mathbb{E}_\theta\left[\nabla_\theta l(X,\theta)\nabla'_\theta l(X,\theta)\right]$ exists, is continuous in θ 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.*

Since in what follows we are considering a drifting sequence of parameters, Assumptions A and B are assumed to hold for each member of this sequence for n large enough.

Lemma 3.1 *Consider the divergences* $d_H(g,\theta_0) = (\theta_0 - \theta_0^c)' \, I(\theta_0) \, (\theta_0 - \theta_0^c)$ *and* $d_W(g,\theta_0) = g'(\theta_0) \left[\nabla'_\theta g(\theta_0) I^{-1}(\theta_0) \nabla_\theta g(\theta_0)\right]^{-1} g(\theta_0)$. *Under Assumptions A and B, 2 KLIC,* d_H*, and* d_W *are locally equivalent.*

The previous lemma yields alternative formulations of the testing problem. 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 involves the norm of the difference between the true and pseudo-true values, 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. 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 model equivalence hypothesis focuses on the restrictions themselves, have a clear intuitive content, and provides further insight on the choice of the tolerance. Relying on this formulation, we can interpret the model equivalence hypothesis as a region of the parameters values "centered" around the restrictions of interest. For instance, when considering an univariate restriction of the form $\theta_{01} = 0$,

$$K_n^W : \; \left| \frac{\theta_{01}}{\sigma_{01}} \right| < \frac{\delta}{\sqrt{n}} \, ,$$

where σ_{01} is the $\sqrt{n}$ asymptotic standard deviation of $\widehat{\theta}_{01}$. In considering a t-test about a mean, Arrow (1960) argued that the "economically significant difference" should be measured in standard deviations units. It is therefore interesting to note that such a standardization appears naturally in the model equivalence approach. The Wald formulation recasts the equivalence hypothesis in terms of parameter values. However it does not tell us exactly what are these values, because σ_{01} is unknown. Now one can always approximate $\sigma_{01}/\sqrt{n}$ by the standard error s_{01}. I define the "equivalence interval" as $(-\delta s_{01}, \delta s_{01})$, that is the approximation of the equivalence hypothesis based on the standard error. An "equivalence region" is defined similarly for multiple restrictions. As will be seen in the next application, such an equivalence interval or region can yield useful information and guidance for implementation.

Let us conclude this section by looking at another divergence.

Lemma 3.2 *Consider the score divergence $d_S = \mathbb{E}_{\theta_0}\nabla'_\theta l(X, \theta_0^c) I^{-1}(\theta_0) \mathbb{E}_{\theta_0}\nabla_\theta l(X, \theta_0^c)$. If the theoretical likelihood equations $\mathbb{E}_{\theta_0}\nabla_\theta l(X, \theta) = 0$ have a unique solution, then under Assumptions A and B, 2 KLIC and d_S are locally equivalent.*

We need a supplementary assumption to obtain local equivalence of the two divergences, because it could be that $\mathbb{E}_{\theta_0}\nabla'_\theta l(X, \theta_0^c)$ is zero, but θ_0^c is distant from θ_0. This phenomenon, scarcely acknowledged in the econometric literature, can happen whenever the likelihood equations have multiple roots. Reeds (1985) illustrates it for a Cauchy model. Freedman (2007) gives an example for a discrete distribution and points out that this yields inconsistency of the classical score significance test. The nonlinear regression Example 2 of Dominguez and Lobato (2004), where $Y = \theta_0^2 X + \theta_0 X^2 + \varepsilon$, can be recast in a maximum likelihood setup assuming $\varepsilon \sim N(0, \sigma^2)$ to give another illustration. These counter-examples show that in nonlinear models the score divergence may not be adapted. However, it can be used whenever the theoretical likelihood equations have a unique root, as in the standard linear regression model. In that case, the hypotheses are

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

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

The model equivalence hypothesis thus focuses on whether the expected score vector of the restricted model is close to zero in the metric defined by $I^{-1}(\theta_0)$.

3.2 Tests and Applications

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

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

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

$$I_n(\theta) = n^{-1}\sum_{t=1}^{n} \nabla_\theta \ln f(X_t;\theta)\nabla_\theta' \ln f(X_t;\theta)\,,$$

without altering the asymptotic properties of each test. Also, as is usual for the score statistic, the information matrix could be evaluated at $\widehat{\theta}_n^c$. Clearly, not every hypotheses-test pair has the same invariance properties. While the $KLIC$-based LR model equivalence test is invariant to (possibly nonlinear) reparameterizations or transformations of the restrictions, the above three tests are invariant to *linear* transformations of the parameter space only. The Hausman-Wald model equivalence test is invariant to nonlinear transformation of the restrictions, while the Wald model equivalence test is invariant to linear transformations only. Invariance of the score model equivalence test is dependent on how the information matrix is evaluated.

Example 1 (continued): *I computed each of the three alternative test statistics H, W, and S. They agree with the LR statistic up to the tenth decimal. All tests then yield the same conclusion for any* δ^2, *and also* $\delta^2_{\text{inf}}(0.01) = 0$ *for each test. 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* Δ *standard deviation of zero for an arbitrary small tolerance* Δ.

It is interesting to compare our Wald equivalence test with the procedure advocated by Andrews (1989). As done in his paper, assume that the restrictions of interest can be reparameterized as $\theta_1 = \mathbf{0}$, with $\theta = (\theta_1, \theta_2)$ and θ_2 is a nuisance parameter. The procedure can be summed up as follows. First test the restrictions using a standard Wald test, which rejects the restrictions at level α if $W_n = n\widehat{\theta}_1'\widehat{\Sigma}_1^{-1}\widehat{\theta}_1 > c_{1-\alpha,r,0}$, where $\widehat{\Sigma}_1$ estimates the $\sqrt{n}$-asymptotic variance of $\widehat{\theta}_1$. If the test does not reject the restrictions, then evaluate for which values of the parameters the test has power at least $p = 1 - \alpha$. If these values are "close" to fulfill the restrictions, conclude that the restrictions approximately hold. Formally, one should estimate the inverse power function of the significance test in direction η and for $\theta_2 = b$, defined as

$$\pi(\eta, p, b) = \inf \left\{ \|\theta_1\| : \theta \in \Theta, \theta_1 \propto \eta, \theta_2 = b, \Pr\left[W_n > c_{1-\alpha,r,0} \mid \theta\right] \geq p \right\} \times \eta \,.$$

Andrews shows that this inverse power function can be estimated through

$$\Pi(\eta, p, b) = \frac{\lambda_{r,\alpha}(p)}{\sqrt{n}} \left(\eta'\widehat{\Sigma}_1^{-1}\eta\right)^{-1/2} \times \eta \,,$$

for values of $\lambda_{r,\alpha}(p)$ tabulated in Andrews (1989). This estimated inverse power function depends on the value of the nuisance parameter θ_2 through $\widehat{\Sigma}_1$. One hence needs to evaluate it at selected values of the nuisance parameter and for different directions to be able to conclude whether the restrictions approximately hold. By contrast, the equivalence approach accepts that the restrictions almost hold by a direct test of

$$H_n^W : \theta_1'\Sigma_1^{-1}\theta_1 \geq \delta^2/n \quad \text{against} \quad K_n^W : \theta_1'\Sigma_1^{-1}\theta_1 < \delta^2/n \,,$$

that (asymptotically) controls $\Pr\left[\text{Reject } H_n^W | H_n^W \text{ true }\right]$, the probability of Type-I error. The two procedures may reach the same qualitative conclusion, as the following example shows, though the details may not coincide.

Example 3 : Exogeneity Restrictions. *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 Andrews, I focus on the "Rate B all customers" results. Lillard and Aigner found that the correlation coefficients are jointly insignificant at the 5% level using a LR significance test. Andrews argued that this conclusion does not seem warranted, as the estimated inverse power measures of the two univariate significance Wald tests indicate that correlations of* $|\rho_1| = 0.47$ *and* $|\rho_2| = 0.55$ *have 50% chances of going undetected.*

For the model equivalence test, the LR statistic equals 1.8 *and* $\delta^2_{\text{inf}}(5\%) = 7.37$. *This value could be used to evaluate the joint "equivalence region" for* $\rho_{k\varepsilon}$ *and* $\rho_{r\varepsilon}$, *were the full covariance matrix of the parameters provided in the original paper. Since it is not, I instead based my reasoning on univariate Wald equivalence tests. Calculations are first based on the parameterization used by Lillard and Aigner, viz.,* $\alpha_{k\varepsilon} = \tan(\rho_{k\varepsilon}\pi/2)$ *and* $\alpha_{r\varepsilon} = \tan(\rho_{r\varepsilon}\pi/2)$, *and then translated into correlation terms. For* $\alpha_{k\varepsilon}$ *and* $\alpha_{r\varepsilon}$, $\delta_{\text{inf}}(5\%) = 3.051$ *and* 2.88 *respectively. These give the equivalence intervals* $|\alpha_{k\varepsilon}| < 1.414$ *and* $|\alpha_{r\varepsilon}| < 1.718$.[8] *The corresponding equivalence intervals for correlations are* $|\rho_{k\varepsilon}| < 0.714$ *and* $|\rho_{r\varepsilon}| < 0.731$. *It is highly unlikely that one would consider such large correlations as evidence of exogeneity. Model equivalence testing thus does not allow to conclude that exogeneity holds, even in an approximate sense.*

4 Asymptotic Properties

I now turn to the formal properties of the tests. It is well known that in general there is no asymptotically uniformly most powerful (UMP) tests in parametric models, so that it is necessary to adopt a local approach in the search of optimal tests, see e.g. Lehmann and Romano (2005). I adopt such an approach and I focus on two criteria for evaluating the model equivalence tests. The first is the local asymptotic maximin criterion, which is also used to characterize the classical trinity of significance tests in parametric models with multivariate parameters, see Borovkov (1998) and Lehmann and Romano (2005). The latters note that the maximin approach may not be compelling for multiparameter significance hypotheses because the distant hypothesis can be defined through different

[8] Standard errors for $\widehat{\alpha}_{k\varepsilon}$ and $\alpha_{r\varepsilon}$ are 0.4635 and 0.5966 respectively.

norms. In the model equivalence framework however, the form of the distant hypothesis is dictated by the considered hypotheses. I found that the model equivalence tests are locally asymptotically maximin, and as a consequence are locally asymptotically unbiased and most powerful against $g(\theta_0) = 0$. The second criterion is local power in the class of tests invariant to orthogonal transformations. Asymptotic invariance to linear transformations is considered by Choi, Hall, and Schick (1998) to show optimality of classical two-sided significance tests of multivariate parameters. I found that model equivalence tests are locally asymptotically UMP among tests invariant to orthogonal transformations of the parameter space, which is a mild requirement fulfilled even by the Wald test. Moreover, 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 the case. Existing results on significance tests use either a restriction that completely determines the parameter value, see e.g. Lehmann and Romano (2005), or a nuisance parameter approach, see e.g. Choi, Hall, and Schick (1998). Specifically, one considers a (possibly nonlinear) reparametrization such that $\theta_0 = (\theta_{01}, \theta_{02})$ and the restrictions completely determine the value of θ_{01}. The component θ_{02}, which is unconstrained under the null hypothesis, is then treated as a nuisance parameter. The formal analysis is based on the score test and the "effective score," that basically purges the score from the nuisance parameter's influence. Such an approach is not suitable for model equivalence testing. First, the restrictions are not assumed to hold under our null hypothesis. Second, and as a consequence, the model equivalence hypothesis generally involve the whole parameter vector, even when the restrictions of interest concern only a subset of them. Third, the score equivalence test is valid only under the restrictive assumption of a unique root for the theoretical likelihood equations, as explained in the previous section. As a result, the theoretical analysis of model equivalence tests cannot rely on the efficient score approach. Our analysis cannot either directly extend Romano's approach (2004), because the latter relies on the univariate dimension of the restriction where $g(\cdot)$ can take positive as well as negative values.

I give here results for the four equivalence tests in a compact way, but it should be understood that each test is considered in turn for testing the corresponding hypotheses, as spelled out in Section 3. For any $\gamma^2 < \delta^2$, define $K_n^{LR}(\gamma) = \{\theta_0 : 2\ KLIC \leq \gamma^2/n\}$ and its boundary $\partial K_n^{LR}(\gamma) = \{\theta_0 : 2\ KLIC = \gamma^2/n\}$. For $J = H, W$, or S, define $K_n^J(\gamma)$ and $\partial K_n^J(\gamma)$ as the similar sets based on the different divergence measures defined in the previous section. The introduction of $K_n^J(\gamma)$ allows to focus on alternatives distant from the null hypothesis, as is usual for maximin analysis.

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. Let J be LR, H, W, or S. If $J = S$, assume that the theoretical likelihood equations have a unique root.*
(A) Let φ_n be a pointwise asymptotically level α tests sequence, that is

$$\limsup_{n\to\infty} \mathbb{E}_\theta \varphi_n \leq \alpha \quad \forall\ \theta \in H_n^J\ .$$

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 $\gamma^2 < \delta^2$,*

$$\limsup_{n\to\infty} \inf_{\theta_0 \in K_n^J(\gamma) \cap \mathcal{N}(\bar\theta, M)} \mathbb{E}_{\theta_0} \varphi_n \leq \Pr\left[\chi_r^2(\gamma^2) < c_{\alpha,r,\delta^2}\right]\ . \tag{4.3}$$

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

$$\limsup_{n\to\infty} \mathbb{E}_{\theta_0} \varphi_n \leq \Pr\left[\chi_r^2(\gamma^2) < c_{\alpha,r,\delta^2}\right]\ . \tag{4.4}$$

(B) The tests sequence π_n^J

1. *is pointwise asymptotically level α,*

2. *is locally asymptotically maximin, in the sense that Inequality (4.3) is an equality for π_n^J, and as a consequence, is locally asymptotically unbiased and most powerful against $\theta_0 = \bar\theta$.*

3. *is locally asymptotically UMP among tests invariant to orthogonal transformations, i.e. Inequality (4.4) is an equality for* π_n^J.

In the formal analysis, I 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 the normal experiment when we observe a sample of size one from $Z \sim N(\mu, \Sigma)$ and we 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, I state here the result that characterizes the UMP invariant test for this problem.

Lemma 4.2 *Consider testing* H *against* K *from one observation* z *from* $Z \in \mathbb{R}^p$ *that follows a multivariate normal* $N(\mu, \Sigma)$ *with unknown mean* μ *and known nonsingular covariance matrix* Σ*. Then the test* $\pi(z)$ *that rejects* H *when* $z'\Sigma^{-1/2}P\Sigma^{-1/2}z < c_{\alpha,r,\delta^2}$ *is of level* α*. For any* $\gamma^2 < \delta^2$*, the test is maximin among* α*-level tests against* $K(\gamma)$: $\mu'\Sigma^{-1/2}P\Sigma^{-1/2}\mu \leq \gamma^2$ *with guaranteed power* $\Pr\left[\chi_r^2(\gamma^2) < c_{\alpha,r,\delta^2}\right]$.

Since the test $\pi(z)$ is maximin, it is necessarily admissible and unbiased. Moreover, as it is independent of γ^2, it must be most powerful against $\mu = 0$. Finally, as it is also invariant to orthogonal transformations of the parameter space, it must be UMP invariant. These properties yield equivalent local asymptotic properties for the model equivalence tests.

I 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, then our Wald approximate hypotheses write

$$H_n : \ |g(\theta_0)| \geq \sigma\delta/\sqrt{n} \qquad \text{against} \qquad K_n : \ |g(\theta_0)| < \sigma\delta/\sqrt{n} \ ,$$

where $\sigma^2 = \nabla'_\theta g(\theta_0) I^{-1}(\theta_0) \nabla_\theta g(\theta_0)$. Romano (2004) considered testing

$$\tilde{H}_n : \ |g(\theta)| \geq \tilde{\delta}/\sqrt{n} \qquad \text{against} \qquad \tilde{K}_n : \ |g(\theta)| < \tilde{\delta}/\sqrt{n} \ .$$

Clearly, the two set of hypotheses are equivalent for $\tilde{\delta} = \sigma\delta$. Romano's test π_n^R rejects $\tilde{H}_n$ if $n^{1/2}|g(\widehat{\theta}_n)| < C(\alpha, \tilde{\delta}, \widehat{\sigma}_n)$, where $\widehat{\sigma}_n^2 = \nabla'_\theta g(\widehat{\theta}_n) I^{-1}(\widehat{\theta}_n)\nabla_\theta g(\widehat{\theta}_n) = \sigma^2 + o_p(1)$ and $C = C(\alpha, \delta, \sigma)$ is the solution of

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

with $\Phi(\cdot)$ the c.d.f. of a $N(0,1)$. Now,

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

see Equation (6) in Romano (2005). Since $C(\alpha, \delta, 1)$ 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 that rejects H_n if

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

It is clear that $C^2(\alpha, \delta, 1) = c_{\alpha,1,\delta^2}$, so π_n^R is asymptotically equivalent to our Wald model equivalence test. Since the other model equivalence tests are also asymptotically equivalent to the Wald model equivalence test, the local asymptotic UMP property of Romano's test (2005, Theorem 3.1) extends to each model equivalence tests. The above reasoning allows to state the following result.

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

$$\limsup_{n\to\infty} \mathbb{E}_\theta \varphi_n \leq \alpha \quad \forall\, \theta \in H_n^J\,.$$

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

$$\limsup_{n\to\infty} \mathbb{E}_{\theta_0} \varphi_n \leq \Pr\left[\chi_1^2(\gamma^2) < c_{\alpha,1,\delta^2}\right]\,. \tag{4.5}$$

Moreover, the tests sequence π_n^J is pointwise asymptotically level α and is locally asymptotically UMP, i.e. Inequality (4.5) is an equality for π_n^J.

5 Conclusion

I 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, as is the standard likelihood ratio significance test. The model equivalence hypothesis under test states that the divergence between the restricted and unrestricted model is smaller than some small tolerance. I also investigated alternative formulation of this hypothesis. The likelihood-ratio model equivalence test, as well as its variants derived from alternative formulations of the hypotheses, have desirable optimality properties. Moreover I have shown through three examples that these tests are easy to apply and can prove useful in practical applications.

I focused on purpose on a well specified parametric model, i.e. that contains the true data generating process, while the restrictions are not supposed to perfectly hold. This allowed us to obtain pretty strong theoretical results. Clearly one would like to extend model equivalence tests to more general contexts where the parametric model could be misspecified or to semiparametric models. The latter would allow to propose model equivalence tests for overidentifying restrictions. The theoretical derivation and practical properties of such tests will be explored in future research.

6 Proofs

In the proofs, I consider drifting sequences of parameters $\{\theta_{0n}, n \geq 1\}$, together with the corresponding sequence $\{\theta_{0n}^c, n \geq 1\}$ defined through (2.1), but for the sake of convenience, the indexes n are omitted throughout. I also omit arguments for the divergences, so that I simply write $KLIC$ instead of $KLIC(g, \theta_0)$.

Proof of Lemmas 3.1 and 3.2. Let us first consider the case where $KLIC = o(1)$. By Assumption A and the information inequality, $\mathbb{E}_{\theta_0} l(X, \theta)$ is continuous in θ and attains its unique maximum at θ_0. Hence $2\ KLIC = o(1)$ implies $\|\theta_0 - \theta_0^c\| = o(1)$. From a Taylor expansion, the information matrix equality, and the continuity of $I(\theta)$ around θ_0,

$$2\ KLIC = (\theta_0 - \theta_0^c)' \, I(\theta_0) \, (\theta_0 - \theta_0^c) \, (1 + o(1)) = d_H \, (1 + o(1)) \ . \tag{6.6}$$

Now use

$$\mathbb{E}_{\theta_0}\nabla_\theta l(X,\theta_0^c) = I(\theta_0)\left(\theta_0-\theta_0^c\right)\left(1+o(1)\right)\ , \tag{6.7}$$

to show that $d_S = d_H\left(1+o(1)\right)$. From Assumption B,

$$0 = g(\theta_0^c) = g(\theta_0) + \nabla'_\theta g(\theta_0)\left(\theta_0-\theta_0^c\right)\left(1+o(1)\right)\ .$$

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

$$\begin{aligned} d_W &= g'(\theta_0)\left[\nabla'_\theta g(\theta_0)I^{-1}(\theta_0)\nabla_\theta g(\theta_0)\right]^{-1}g(\theta_0) \\ &= \left(\theta_0-\theta_0^c\right)'I^{1/2}(\theta_0)P_0I^{1/2}(\theta_0)\left(\theta_0-\theta_0^c\right)'\left(1+o(1)\right)\ . \end{aligned} \tag{6.8}$$

The constrained optimization problem for θ_0^c yields $\mathbb{E}_{\theta_0}\nabla_\theta l(X,\theta_0^c) = \nabla_\theta g(\theta_0^c)\lambda$ for some $\lambda\in\mathbb{R}^r$. From $\|\theta_0-\theta_0^c\| = o(1)$ and (6.7), $I^{-1/2}(\theta_0)\nabla_\theta g(\theta_0^c)\lambda = I^{1/2}(\theta_0)\left(\theta_0-\theta_0^c\right)\left(1+o(1)\right)$. From Assumption B, $\nabla_\theta g(\theta_0^c) = \nabla_\theta g(\theta_0)\left(1+o(1)\right)$ and both matrices have the same rank. Combine these facts to obtain $P_0I^{1/2}(\theta_0)\left(\theta_0-\theta_0^c\right) = I^{1/2}(\theta_0)\left(\theta_0-\theta_0^c\right)\left(1+o(1)\right)$, so that from (6.8) $d_W = d_H(1+o(1))$.

If $d_H = o(1)$, then because $I(\theta_0)$ is non-singular, $\|\theta_0-\theta_0^c\| = o(1)$, and (6.6) yields local equivalence with $2\ KLIC$ and d_W.

If $d_W = o(1)$, because $I(\theta_0)$ is non-singular and $\nabla_\theta g(\theta_0)$ is full rank, $g(\theta_0) = o(1)$. By Assumption B there exists $\bar\theta$ such that $\|\theta_0-\bar\theta\| = o(1)$ with $g(\bar\theta) = 0$. Therefore

$$0 \le 2\ KLIC \le 2\ \mathbb{E}_{\theta_0}\frac{l(X,\theta_0)}{l(X,\bar\theta)} = \left(\theta_0-\bar\theta\right)'I(\theta_0)\left(\theta_0-\bar\theta\right)\left(1+o(1)\right) = o(1)\,.$$

But $2\ KLIC = o(1)$ implies local equivalence with divergences d_H and d_W as shown above.

Finally, if $d_S = o(1)$, then under our supplementary assumption, (6.7) holds, and local equivalence with d_H and other divergences follows. ■

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

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

Moreover, $\mu' P \mu = \mu' A A' P A A' \mu = \varepsilon_r' \varepsilon_r$, 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 maximin 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 last equality holds by a rotation of the space that makes ε_{r1} parallel to e_x while leaving Q_1 invariant. The rejection region of the test is thus

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

for some constant $A > 0$. The function $\psi(\cdot)$ is positive and strictly increasing on $(0, +\infty)$ with $\psi(0) = 1$ and $\psi'(0) = 0$. It is also logarithmically strictly convex. Indeed, by Holder's inequality, for $t \neq u$ and $0 < \lambda < 1$,

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

Hence, $h'(0) = 0$ and $h'(t) < 0$ for all $t > 0$. So there is at most one t_0 such that $h(t_0) = 0$, and there should be at least one such t_0 for a test with level $0 < \alpha < 1$. Therefore t_0 is unique, and the test is $\|x_r\|^2 < c$ for some constant c, which also writes

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

The most powerful Bayes test of level α obtains for $c = c_{\alpha,r,\delta^2}$. Let us check that this test is maximin 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 maximin, see e.g. Borovkov (1998, Theorem 49.1), and is unbiased by definition of a maximin 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. I focus on the Hausman-Wald test, which is more convenient to deal with because it involves the basic parameter vector. I 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 $\mathbb{E}_{\theta_0} l(X, \theta_0) \geq \mathbb{E}_{\theta_0} l(X, \theta_0^c) \geq \mathbb{E}_{\theta_0} l(X, \bar{\theta})$,

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

Since $\|\theta_0 - \bar{\theta}\| = O(n^{-1/2})$ for $\theta_0 \in \mathcal{N}(\bar{\theta}, M)$ by the definition of $\mathcal{N}(\bar{\theta}, M)$, a Taylor expansion yields

$$\mathbb{E}_{\theta_0} l(X, \bar{\theta}) - \mathbb{E}_{\theta_0} l(X, \theta_0) = (1/2)\left(\theta_0 - \bar{\theta}\right)' I(\theta_0) \left(\theta_0 - \bar{\theta}\right)(1 + o(1)) = O(n^{-1}) \qquad (6.10)$$

uniformly in $\theta_0 \in \mathcal{N}(\bar{\theta}, M)$. Therefore, from (6.9), $KLIC = O(n^{-1})$, and thus $d_H = O(n^{-1})$ by Lemma 3.1, so that $\|\theta_0 - \theta_0^c\| = O(n^{-1/2})$. From Lemma 3.1'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 $\mathcal{N}(\bar{\theta}, M)$,

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

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

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

Expand the right-hand side term of (6.11) to obtain that uniformly in $\theta_0 \in \mathcal{N}(\bar{\theta}, M)$

$$(\theta_0 - \theta_0^c)' \, I_0 \, (\theta_0 - \theta_0^c)' = n^{-1} h' \bar{I}^{1/2} \bar{P} \bar{I}^{1/2} h + o(n^{-1/2}) \, . \qquad (6.12)$$

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 $h'\bar{I}^{1/2}\bar{P}\bar{I}^{1/2}h < \delta^2$, apply Lemma 4.2 to deduce the bounds (4.3) and (4.4).

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 $\mathbb{P}^n_{\bar{\theta}}$

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

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 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. Moreover, for θ_0 such that $n^{1/2}\min_{g(\theta)=0}\|\theta_0 - \theta\| \to \infty$, $|\tau_n| \to \infty$ and the power of both tests tends pointwise to zero.

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,

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

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.

Since the power of $\pi(\tau_n)$ converges to a bounded function continuous in θ_0, limits of extrema on $K(\gamma)$ equal limits of extrema on $K_n^H(\gamma)$ using (6.12). Hence the same local asymptotic properties hold for $\pi(\tau_n)$, and thus π_n^H, as tests of H_n^H against $K_n^H(\gamma)$.

v. To extend the result to the LR test, use a similar reasoning and (6.10) to deduce that limits of extrema on $K_n^H \cap \mathcal{N}(\bar{\theta}, M)$ equal limits of extrema on $K_n^{LR} \cap \mathcal{N}(\bar{\theta}, M)$. 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 Wald and Score 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(5), 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 University Press: Stanford.

Berger, J.O., and Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconciliability of P Values and Evidence. (with comments and rejoinder) *J. Amer. Statist. Assoc.* 82(397), 112–139.

Berger, J.O., and Delampady, M. (1987). Testing Precise Hypotheses. (with comments and rejoinder) *Statis. Sci.* 2(3), 317–352.

Berkson, J. (1938). Some Difficulties on Interpretation Encountered in the Application of the Chi-Square Test. *J. Amer. Statist. Assoc.* 33(203), 526–536.

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.

Casella, G. and Berger R.L. (1987). Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. *J. Amer. Statist. Assoc.* 82(397), 106–111.

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). Some Problems Connected with Statistical Inference. *Annals of Math. Statist.* 29(2), 357–372.

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.

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

Freedman, D.A. (2007). How Can the Score Test Be Inconsistent? *Amer. Statist.* 61(4), 291–295.

Godfrey, L.G. (1988). *Misspecification Test in Econometrics.* Cambridge University Press: New-York.

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(6), 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.

Nelson, F.D., and Savin, N.E. (1990). The Danger of Extrapolating Asymptotic Local Power. *Econometrica* 58(4), 977–981.

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

Reeds, J.A. (1985). Asymptotic Number of Roots of Cauchy Location Likelihood Equations. *Ann. Statist.* 13(2), 775–784.

Romano, J.P. (2005). Optimal Testing of Equivalence Hypotheses. *Ann. Statist.* 33(3), 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.

SENN, 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.

Table 1: Critical Values for r=1 (10% and 5% level)

δ^2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.017	0.019	0.021	0.024	0.026	0.029	0.032	0.035	0.039	0.043
	0.004	0.005	0.005	0.006	0.006	0.007	0.008	0.009	0.010	0.011
1	0.047	0.052	0.057	0.063	0.070	0.077	0.084	0.093	0.102	0.112
	0.012	0.013	0.014	0.016	0.018	0.019	0.021	0.024	0.026	0.029
2	0.123	0.135	0.147	0.161	0.176	0.192	0.209	0.227	0.247	0.268
	0.032	0.035	0.039	0.042	0.047	0.051	0.057	0.062	0.068	0.075
3	0.290	0.313	0.337	0.363	0.390	0.418	0.447	0.478	0.509	0.542
	0.082	0.090	0.099	0.108	0.118	0.129	0.141	0.153	0.167	0.181
4	0.575	0.610	0.646	0.682	0.719	0.758	0.797	0.837	0.877	0.919
	0.196	0.212	0.229	0.247	0.266	0.286	0.307	0.328	0.351	0.375

Table 2: Critical Values for r=2 (10% and 5% level)

δ^2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.222	0.233	0.245	0.257	0.270	0.284	0.298	0.313	0.328	0.344
	0.108	0.113	0.119	0.125	0.132	0.138	0.145	0.153	0.160	0.168
1	0.361	0.379	0.397	0.416	0.436	0.457	0.478	0.500	0.523	0.547
	0.177	0.186	0.195	0.205	0.215	0.225	0.236	0.248	0.260	0.272
2	0.571	0.597	0.623	0.650	0.678	0.707	0.736	0.766	0.798	0.830
	0.285	0.299	0.313	0.328	0.343	0.359	0.375	0.392	0.410	0.428
3	0.862	0.896	0.930	0.965	1.001	1.037	1.074	1.112	1.151	1.190
	0.447	0.466	0.487	0.507	0.529	0.551	0.574	0.597	0.621	0.646
4	1.230	1.271	1.312	1.353	1.396	1.439	1.482	1.526	1.571	1.616
	0.671	0.697	0.723	0.751	0.778	0.807	0.836	0.865	0.895	0.926

Table 3: Critical Values for r=3 (10% and 5% level)

δ^2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.604	0.624	0.645	0.667	0.689	0.712	0.735	0.759	0.784	0.809
	0.364	0.376	0.389	0.402	0.415	0.429	0.443	0.458	0.473	0.489
1	0.835	0.862	0.889	0.917	0.946	0.975	1.005	1.035	1.066	1.098
	0.505	0.521	0.538	0.555	0.573	0.592	0.611	0.630	0.650	0.670
2	1.131	1.164	1.198	1.232	1.267	1.303	1.339	1.376	1.413	1.452
	0.691	0.712	0.734	0.756	0.779	0.803	0.826	0.851	0.876	0.901
3	1.490	1.530	1.569	1.610	1.651	1.692	1.734	1.777	1.820	1.864
	0.927	0.954	0.981	1.008	1.036	1.065	1.094	1.124	1.154	1.184
4	1.908	1.952	1.997	2.043	2.089	2.135	2.182	2.229	2.277	2.325
	1.215	1.247	1.279	1.311	1.344	1.378	1.412	1.446	1.481	1.516

Table 4: Critical Values for r=4 (10% and 5% level)

δ^2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	1.090	1.118	1.146	1.174	1.203	1.233	1.263	1.294	1.325	1.357
	0.729	0.747	0.766	0.785	0.805	0.825	0.845	0.866	0.887	0.909
1	1.390	1.423	1.456	1.490	1.525	1.560	1.596	1.632	1.669	1.707
	0.931	0.953	0.976	1.000	1.024	1.048	1.073	1.098	1.124	1.150
2	1.745	1.783	1.822	1.861	1.901	1.942	1.983	2.024	2.066	2.109
	1.176	1.203	1.231	1.259	1.287	1.316	1.345	1.375	1.405	1.436
3	2.152	2.195	2.239	2.283	2.328	2.373	2.419	2.465	2.511	2.558
	1.467	1.498	1.530	1.562	1.595	1.628	1.662	1.696	1.730	1.765
4	2.606	2.653	2.701	2.750	2.799	2.848	2.898	2.947	2.998	3.048
	1.800	1.836	1.872	1.908	1.945	1.983	2.020	2.058	2.096	2.135

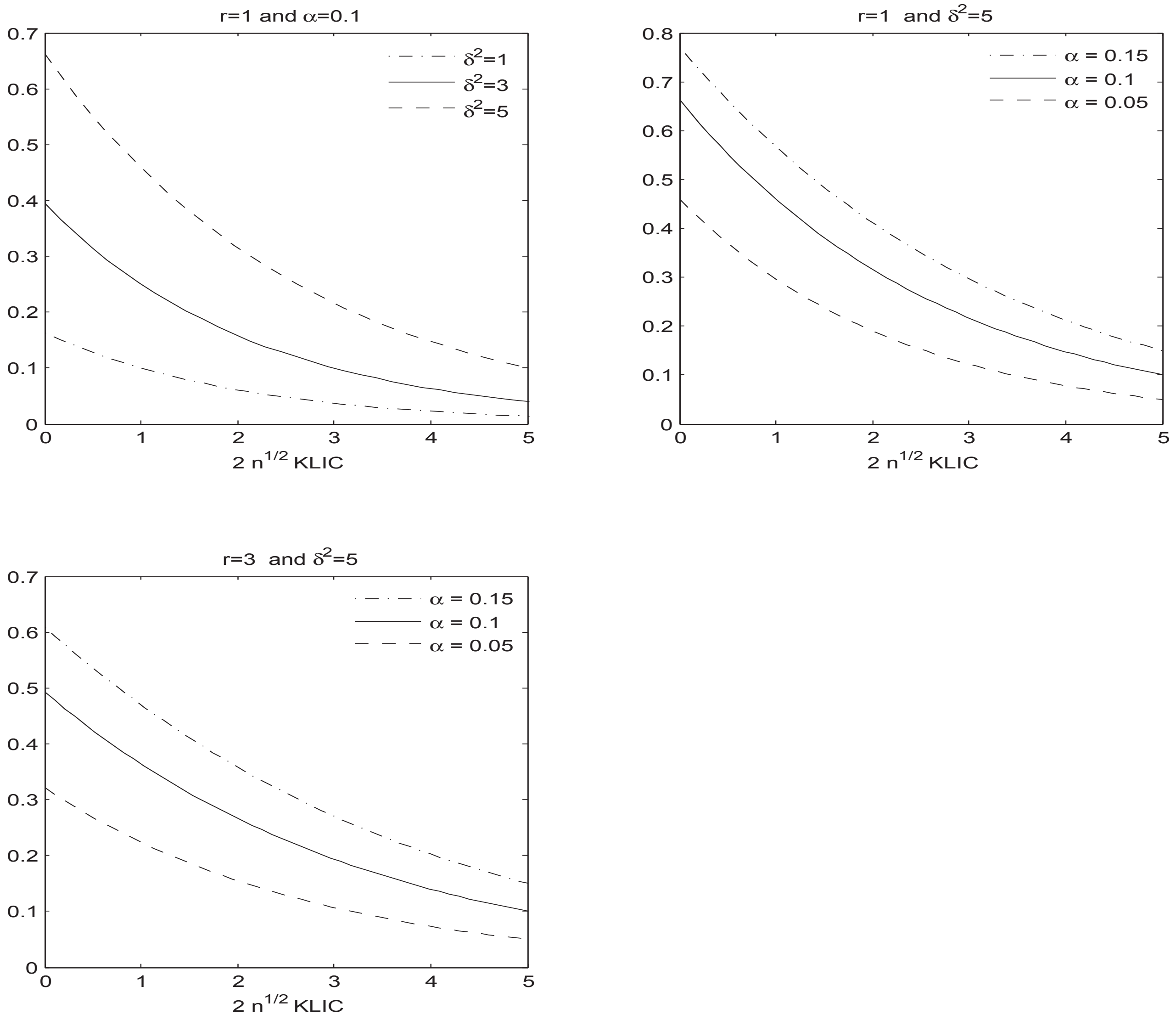
Table 5: Critical Values for r=5 (10% and 5% level)

δ^2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	1.643	1.676	1.709	1.743	1.778	1.812	1.848	1.884	1.920	1.957
	1.169	1.192	1.216	1.240	1.265	1.290	1.315	1.341	1.367	1.394
1	1.994	2.032	2.071	2.109	2.149	2.188	2.229	2.269	2.310	2.352
	1.421	1.449	1.476	1.505	1.533	1.562	1.592	1.622	1.652	1.682
2	2.394	2.436	2.479	2.523	2.566	2.611	2.655	2.700	2.746	2.792
	1.713	1.745	1.777	1.809	1.841	1.874	1.908	1.942	1.976	2.010
3	2.838	2.884	2.931	2.979	3.027	3.075	3.123	3.172	3.221	3.271
	2.045	2.081	2.116	2.152	2.189	2.225	2.262	2.300	2.338	2.376
4	3.321	3.371	3.421	3.472	3.524	3.575	3.627	3.679	3.732	3.784
	2.415	2.453	2.493	2.532	2.572	2.612	2.653	2.694	2.735	2.777

Table 6: Critical Values for r=6 (10% and 5% level)

δ^2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	2.241	2.278	2.316	2.355	2.393	2.432	2.472	2.512	2.552	2.593
	1.663	1.691	1.719	1.747	1.776	1.805	1.835	1.865	1.895	1.926
1	2.635	2.676	2.718	2.761	2.804	2.847	2.891	2.935	2.980	3.025
	1.957	1.989	2.020	2.053	2.085	2.118	2.151	2.185	2.219	2.253
2	3.070	3.116	3.162	3.208	3.255	3.302	3.349	3.397	3.446	3.494
	2.288	2.323	2.358	2.394	2.430	2.466	2.503	2.540	2.577	2.615
3	3.543	3.592	3.642	3.692	3.742	3.793	3.843	3.895	3.946	3.998
	2.653	2.692	2.730	2.769	2.809	2.849	2.889	2.929	2.970	3.011
4	4.050	4.102	4.155	4.208	4.261	4.315	4.369	4.423	4.477	4.532
	3.052	3.093	3.135	3.177	3.220	3.263	3.306	3.349	3.393	3.436

Figure 1: Asymptotic power curves


r=1 and α=0.1
δ²=1
δ²=3
δ²=5
2 n^{1/2} KLIC
r=1 and δ²=5
α = 0.15
α = 0.1
α = 0.05
2 n^{1/2} KLIC
r=3 and δ²=5
α = 0.15
α = 0.1
α = 0.05
2 n^{1/2} KLIC