

# Hypothesis Testing and the Search for Global Theories

Scott Gilbert  
Southern Illinois University, Carbondale  
gilberts@siu.edu

February 2004

## Abstract

An important objective of science is to find global theories, those that explain/predict what happens in a wide variety of circumstances. Along the way, scientists usually encounter local theories which are either discarded or embedded in a more general theory. Statistical hypothesis tests provide two tools for this scientific method: (a) Tests for theory significance, regardless of local/global distinction, and (b) Tests for global-ness versus local-ness. The present work takes pieces of information from each method and builds some new tests, with power focused on global theories. The tests answer the question: “Is the theory valid and global?”, rather than a subordinate question: “Is it valid?” or “Is it global?”. The statistics are asymptotically equivalent to quadratic forms in statistics obtained from standard methods (a) and (b), and under simplifying assumptions these forms coincide with out-of-sample and nested-sample model validation statistics. We examine test performance in simulation, and illustrate with an economic example.

Keywords: Hypothesis test, global, local, parameter change, in-sample, out-of-sample, nested-sample.

A previous version of this paper was presented at the Midwest Econometrics Group, the University of Mississippi, and Southern Illinois University - Carbondale, and for helpful comments I would like to thank seminar participants including Michael McCracken, Todd Clark, Richard Ashley, Barbara Rossi, Walter Enders, Shinishi Sakata, Michael Belongia, Walter Mayer, Mark VanBoenig, John Conlon, Zsolt Besci, Sajal Lahiri and Kevin Sylwester.

## 1. INTRODUCTION

An important objective of science is to find global theories, those that explain/predict what happens in a wide variety of circumstances. Along the way, scientists usually encounter local theories which are either discarded or are embedded in a more general theory, and the frontier of many sciences can be defined by the most recent efforts to convert/combine local theories into global ones. For example, in economics the last two decades have seen new attempts to find a theory that fits both poor and rich countries, those with capitalism, communism, freedom, repression, etc. These attempts have led both to a “new growth theory” and a closer look at the role of institutions in determining economic outcomes (see Romer 2001 for an overview). In the natural sciences, the physics community has labored for over a century to build a global theory of energy, matter and motion. Biology, with its genome projects, is now able to identify those parts of the genetic code which are “global” for a broad group of organisms, such as the primates.

The development of global scientific theories requires enormous effort, and to assist this process statisticians have invented useful testing procedures. Some of these procedures check the overall significance or explanatory power of a theory. Others check whether or not a theory is global (rather than local) in scope, being equally applicable to all parts of the relevant population. Hence, statistical hypothesis tests provide two tools for the scientific method:

- (a) Tests for theory significance, regardless of local/global distinction.
- (b) Tests for global-ness versus local-ness.

In principle, significance tests (a) are to be applied only to samples from a homogeneous population, thereby avoiding the problem of local-ness caused by population heterogeneity; however, the ultimate aim of scientific theories is to explain as much behavior as possible, causing frequent application to rather broad datasets.

To keep things simple, suppose that a theory applies to a population composed of just two parts, and that to test the theory we have a sample of  $i = 1, 2, \dots, m, m + 1, m + 2, \dots, n$  observations, with observations  $1, 2, \dots, m$  drawn from the first sub-population, and observations  $m + 1, m + 2, \dots, n$  drawn from the second. There are then  $n_1 = m$  observations in the first sample,  $n_2 = n - m$  observations in the second one. Theories that explain/predict a variable  $y$ , given another variable or control  $x$ , are often applied via a linear model (which we use here only as an example):

$$y = \alpha + \beta x + \varepsilon,$$

with error  $\varepsilon$  normally distributed  $N(0, \sigma^2)$ , independent of  $x$ . In this setting, the theory has explanatory power if coefficient value  $\beta \neq 0$  is the best choice of  $\beta$ , at least for some part of the population, and a theory is global if there is a single best choice of  $\beta$  for each and every part of the population. A standard  $t$  test of the hypothesis  $H_0 : \beta = 0$ , when applied to the whole sample  $i = 1, 2, \dots, n$ , reports on the overall significance of the theory. A two-sample test for parameter equality, across sub-populations, can be performed via the augmented regression  $y = \alpha + \beta x + \gamma D x + \varepsilon$ , where  $D$  is a dummy variable = 1 for the first sub-population, = 0 otherwise. The test (sometimes called a Chow test, after Chow 1960) can be based on the  $t$  statistic for  $\gamma$  in this regression. If we want to find out whether the theory is global *and* significant, we can use the two tests together somehow. Most commonly, we can first apply the two-sample test of equality in parameter values, then the test for parameter significance, and if the result is (“fail to reject”, “reject”) then the global+significant view is deemed credible. This approach has some limitations (see below) which partly motivate our proposed new tests.

Abstracting to a more general class of models (such as multivariate linear models, non-linear models, etc.), suppose that a theory has explanatory power if, for some  $p \times 1$  parameter vector  $\theta$  (and an integer  $p \geq 1$ ), the overall

best choice of  $\theta$  has some non-zero elements, and that the theory is global if this choice is the same across sub-populations. In the above example,  $p = 1$  and  $\theta = \beta$ . The null hypothesis is  $H_0: \theta = 0$ , meaning that all elements of  $\theta$  equal 0. Let  $U$  be a non-negative statistic for an upper-tailed test of significance for  $\theta$ , such that  $U$  has an asymptotic (large-sample) chi square distribution, with  $p$  degrees of freedom, under  $H_0$ . An example is  $t^2$  (with  $t$  a student's  $t$  statistic) or more generally  $pF$ , with Fisher's  $F$  statistic. Let  $W$  be a non-negative statistic for an upper-tailed two-sample test of differences in  $\theta$  across two (exclusive, exhaustive) sub-populations, distributed chi square ( $p$  degrees of freedom) asymptotically under  $H_0$ , independent of  $U$ . For example,  $W$  could be a squared  $t$  statistic for a two-sample (Chow) test. A joint assessment of significance *and* global-ness can, if desired, be based on a "joint" statistic:  $J = \max(U - c_U, c_W - W)$ , with  $c_U$  and  $c_W$  being the chosen critical values for the individual upper-tailed  $U$  and  $W$  tests. This "joint" test, with rejection rule  $J > 0$ , rejects the null only if parameters  $\theta$  are significant and intra-population parameter difference is insignificant.

On a practical level the joint test  $J$ , if done carefully, has two limitations: (i) If, as is common, the individual tests  $U$  and  $W$  are done at several significance levels (say 10%, 5%, 1%), then there are nine or more versions of the joint test to be reported, and for each of these the joint significance level must be determined, (ii) the joint test's  $p$ -value, indicating the threshold level of significance, is not uniquely defined, and instead depends on the pairing of significance levels for the  $U$  and  $W$  tests.

The present work proposes some new tests for global theories, with some of the virtues of the joint test  $J$ , while avoiding the above-mentioned limitations of  $J$ . The proposed tests are not intended to replace traditional tests (symbolized by  $U$ , above) of significance, but to give a variation on them useful for some purposes. Also, the proposed tests link the problem of theory testing to the practise of out-of-sample or nested-sample model validation,

in a way that adds perspective to both disciplines. A key virtue of  $J$ , in the above example and similar situations, is that  $J$  is more likely to attach significance to parameters  $\theta$  when their values remain constant across the population. The test thereby seeks to answer the question: “Is the theory valid and global?”, rather than a subordinate question: “Is it valid?” or “Is it global?”. To avoid the noted limitations of  $J$ , we propose statistics which are (asymptotically equivalent to) suitably chosen functions of the information in the underlying test statistics  $U$  and  $W$ .

To define the proposed tests we first write  $U = u'u$  and  $W = w'w$ , for some  $p \times 1$  vectors  $u$  and  $w$  which in large samples are assumed to be mutually independent and standard normal under  $H_0$  (see Assumption 2, Section 2). In the above example,  $U = t^2$  so we can set  $u = t$ ; similarly we can let  $w$  be a two-sample  $t$  statistic. The most basic form of the proposed tests statistics is a quadratic function of  $u$  and  $w$ :

$$G = u'u - a u'w - b w'w, \tag{1}$$

for some constants  $a$  and  $b$ ,  $a \geq 0$  and  $b > 0$ , which can depend on sample size but converge to large-sample limits. Equivalently,  $G = U - bW - a u'w$ . The proposed  $G$  test is upper-tailed, and since we are free to scale  $G$  by a (positive) constant,  $G$  effectively includes the more general form  $G = cu'u - au'w - bw'w$ , for constants  $a, b, c$ :  $a \geq 0, b > 0, c > 0$ . These restrictions on  $a, b, c$  characterize the proposed tests, while if we relax these restrictions we obtain some other, known forms of  $G$  including  $G = U$ ,  $G = W$  and  $G = U + W$ , the last of these being a test statistic (recently studied by Rossi 2003) for parameter significance *and/or* intra-population parameter differences. Somewhat more generally, we define a class of statistics  $G^*$  for which:

$$G^* = G(1 + o_p(1)),$$

asymptotically equivalent to a  $G$  statistic, in large samples.

How do we choose the constants  $a$  and  $b$  in the proposed test statistic? The closest analogy to the joint statistic  $J$  is  $G = U - W$ , for which  $a = 0$  and  $b = 1$ . With statistics  $U$  and  $W$  having a joint large-sample distribution in which they are independent chi square (with  $p$  degrees of freedom) variables, we can compute large-sample critical values ( $= 2.0689, 3.1904, 5.9672$  at significance levels  $\alpha = 10\%, 5\%, 1\%$ , when  $p = 1$ , see Table 1) for this  $G$  (and, equivalently,  $G^*$ ). We can, further, find the unique threshold significance level (asymptotics-based  $p$ -value) for  $G$ , at which the test just (marginally) rejects  $H_0$ . This approach relies on asymptotic theory, via large-sample critical values, and while exact finite-sample critical values are sometimes preferred, we leave such exact testing to future work.

To get a broader view of what “good” choices for  $a$  and  $b$  might look like, we perform some out-of-sample and in-and-out-of-sample (or “nested” sample, a concept different from the in-and-out-of-sample approach of Pressnell and Boos 2004) model-fitting exercises (detailed in Section 4). In these exercises, there is a “training” sample that consists of our first sub-sample ( $i = 1, 2, \dots, m$ ) (which therefore takes on a special status, see below) and a “validation” sample that consists of either the remaining sub-sample ( $i = m + 1, m + 2, \dots, n$ ) or the whole sample ( $i = 1, 2, \dots, n$ ). We refer to the former as *out-of-sample* validation, and the latter as *nested-sample* validation, for obvious reasons. The model is estimated on the training sample and then applied/fit to the validation sample. The idea here is very simple: If a theory is valid and global then it should perform reasonably well when validated on a sample which differs (in part) from the training sample, whereas if the theory is valid but purely local then performance on the validation sample should be degraded (due to inconsistent parameter estimates). So cross-validation tends to penalize purely local theories, moreso than does in-sample validation.

We note that in economics it is common practice to evaluate time series models based on out-of-sample performance. Some recent papers include Diebold and Mariano (1995), West (1996), Ashley (1998), McCracken (1999), Gilbert (2001), Clark and McCracken (2001a,b) and Inoue and Killian (2003). A reason for this practice is that the economy appears to change enough over time to cause many popular economic theories to become obsolete/incomplete/local at some point (see for example Clements and Hendry 1999), and this intensifies the search (temporally) consistent theories.

Our nested-sample and split-sample model-fitting exercises suggest two ways to specify (a,b):

$$a = 0, \quad b = \frac{n}{m} - 1. \quad (2)$$

$$a = 2\sqrt{\frac{m}{n-m}}, \quad b = \frac{n}{m} + 1. \quad (3)$$

Here,  $m/n$  is the proportion of the total sample comprised of the first (of two) sub-sample(s). Under (2), the closure of the range of values of (a,b) is  $\{a = 0, b \geq 0\}$ , whereas (3) yields a different closure  $\{a = \frac{2}{\sqrt{b-2}}, b \geq 2\}$ . We can compute  $G$  statistics using (2) or (3), and we will refer to these as the  $G_{\text{nest}}$  and  $G_{\text{split}}$  versions of  $G$ , respectively. In the context of some linear regression models (see Section 5), the test statistic  $G_{\text{split}}$  coincides (up to a scalar multiple) with an “out-of-sample F test” proposed independently by McCracken (1999) and Gilbert (2001), and Inoue and Killian (2003) refer to this  $G_{\text{split}}$  as the Gilbert-McCracken test. In the same context, the nested-sample test statistic  $G_{\text{nest}}$  coincides with a “nested-sample F test” proposed by Gilbert (2001).

We show (Theorem 3) that, under simplifying assumptions, there exist ( $G^*$ ) statistics  $G_{\text{nest}}^\dagger$  and  $G_{\text{split}}^\dagger$  which are asymptotically equivalent to  $G_{\text{nest}}$  and  $G_{\text{split}}$ , respectively, such that  $G_{\text{nest}}^\dagger$  is obtained from a nested-sample likelihood-based model validation, and  $G_{\text{split}}^\dagger$  is obtained from split-sample

validation. To our knowledge the statistics  $G_{\text{nest}}^\dagger$  and  $G_{\text{split}}^\dagger$  have not been proposed before. The tests  $G$  have one advantage over their  $G^\dagger$  counterparts, in that they can easily be made robust to intra-population differences in nuisance parameters (e.g. parameters other than  $\theta$ ) such as error variances  $\sigma^2$ .

The training sample ( $i = 1, 2, \dots, m$ ) could instead be specified as the remaining sub-sample ( $i = m + 1, m + 2, \dots, n$ ), and test versions (2) and (3) permit this by just switching the labelling and element numbering of the two sub-samples. The choice of training sample influences  $G_{\text{nest}}$  and  $G_{\text{split}}$  (and  $G_{\text{nest}}^\dagger, G_{\text{split}}^\dagger$ ) via the ratio  $m/n$ , except when each sub-sample is equal-sized ( $m = n - m$ ). Hence, we typically have two ways of doing the proposed tests, depending on the training sample; we might choose just one of these if one sub-sample has some historical or logical precedence (as in Section 6), but otherwise might combine them somehow or report both (plus Bonferroni bounds or other descriptors of joint significance, an exercise we leave to future work).

How does the “canonical” choice  $(a, b) = (0, 1)$  fit into the frameworks (2) and (3)? It fits only into the “nested-sample” framework (2), in the case of equal sub-sample sizes. We can use this canonical test  $G_{\text{nest}}$  even when sub-sample sizes are unequal, because the asymptotic distribution of  $G$  under  $H_0$  is fixed once we specify  $(a, b)$ , but here the (asymptotic) equivalence to nested-sample validation ( $G_{\text{nest}}^\dagger$ ) breaks down. Also, test power is affected by sub-sample sizes, and our (asymptotic-local) power analysis is restricted to tests with  $(a, b)$  specified in terms of actual sample sizes, via (2) or (3).

To summarize the power of the proposed tests, the (large-sample) power of nested-sample test  $G_{\text{nest}}$  is greater in the absence of intra-population parameter differences than in the presence of it (see Theorem 2). In other words, the test is more likely to reveal global+valid theories than global+local theories, whereas a standard significance test  $U$  (applied to the whole sample

$i = 1, 2, \dots, n$ ) is equally likely to reveal valid+global or valid+local theories. On the other hand, the  $U$  test has higher overall power than  $G_{\text{nest}}$ , hence the advantage of  $G_{\text{nest}}$  is in *discriminating* between alternatives (via power differentials), rather than overall power. By comparison, for the split-sample version of  $G_{\text{split}}$ , power is greater under parameter constancy when  $m/n$  is sufficiently small, but the situation is otherwise mixed.

To further interpret the proposed methods suppose that  $\theta$  is a list of (some) parameters for a probability model of a random vector  $z$ , and let  $U = -2 \ln(\lambda)$  with  $\lambda$  a full-sample likelihood ratio (LR) for the constrained model ( $H_0: \theta = 0$ ) versus the unconstrained (all  $\theta$  values) model, each applied to the whole sample  $i = 1, 2, \dots, n$  (with the same  $\theta$  at all  $i$ ). Then the nested-sample form (2) of  $G$  is a “penalized” LR test statistic with stochastic penalty  $-W \frac{n-m}{m}$  having an asymptotic (large-sample) expectation  $\approx -p \frac{n-m}{m}$  under the null, for asymptotically chi square (with  $p$  degrees of freedom)  $W$ . The split-sample form (3) has a more complex interpretation as a modified LR statistic, with “penalty” term  $-au'w - bw'w$  having asymptotic expectation  $\approx -p \frac{n+m}{m}$  when  $u$  and  $w$  are independent and standard normal under  $H_0$ . By comparison, for constrained and unconstrained models Akaike’s (1973) Information Criterion (AIC) selects the latter if and only if  $U - 2p > 0$ , with non-stochastic penalty  $-2p$  which is equal in expectation, asymptotically, to the nested-sample test’s penalty when  $m = n/3$ , and to the split-sample test’s penalty when  $m = n$  (an extreme at which cross-validation is infeasible). In the nested-sample case, the situation  $m = n/3$  yields statistic  $G_{\text{nest}} = U - 2W$ , somewhat different than the “canonical” form  $G_{\text{nest}} = U - W$ , with more severe penalty for intra-population parameter differences.

Like the proposed versions of the statistic  $G$ , model selection criterion AIC can be motivated via a “cross-validation” model-fitting exercise (Stone 1977, and for discussion see Efron and Tibshirani 1993, Shao 1996, 1997 and McQuarrie and Tsai 1998); however, while for  $G$  the relevant “cross-

validation” is simple (one training and one validation sample), Stone (1977) obtains AIC from leave-one-out cross-validation (CV). For linear regression models Zhang (1993) extends Stone’s (1977) results to obtain equivalence between AIC and CV in which several observations are left out (see also Shao 1993). For these models Wei (1992) shows that the Bayes’ information criterion (BIC) is asymptotically equivalent to a form of cross-validated performance measurement involving a sequence of successively updated training samples and one-period-ahead validation samples.

The remainder of the paper is as follows. Sections 2 and 3 describe test distribution and power, and Section 4 connects the tests to cross-validation methods. Section 5 examines the case of regression models, Section 6 illustrates the methods in an economic example, and Section 7 concludes. An Appendix contains mathematical proofs.

## 2. DISTRIBUTION

To obtain the asymptotic distribution of proposed statistics under  $H_0$ , suppose that the proportion  $m/n$  of observations in the first sub-sample approaches a large-sample limit, as follows:

*Assumption 1:*  $\frac{m}{n} \rightarrow \rho$  as  $n \rightarrow \infty$ , for some  $\rho$  in  $(0, 1)$ .

Next, regarding test statistics  $U = u'u$  and  $W = w'w$  (for testing  $\theta$  explanatory power and intra-population  $\theta$  differences, respectively) we have:

*Assumption 2:* Under  $H_0 : \theta = 0$ , the  $2p \times 1$  vector  $= (u', w)'$  converges in distribution to standard normal.

To justify Assumption 2 consider the common setup (referred to as SETUP later), with  $\hat{\theta}_1$  and  $\hat{\theta}_2$  estimators of  $\theta$  computed on the first and second sub-sample, respectively, asymptotically independent normal vectors with  $\sqrt{n}(\hat{\theta}_1 - \theta_1) \xrightarrow{d} N(0, M_1)$  and  $\sqrt{n}(\hat{\theta}_2 - \theta_2) \xrightarrow{d} N(0, M_2)$ , for some invertible

variance-covariance matrices  $M_1$  and  $M_2$ . Let  $\hat{\theta}$  be a full-sample estimator  $\hat{\theta} \approx (\hat{M}_1^{-1} + \hat{M}_2^{-1})^{-1}(\hat{M}_1^{-1}\hat{\theta}_1 + \hat{M}_2^{-1}\hat{\theta}_2)$  asymptotically, with  $\hat{M}_1$  and  $\hat{M}_2$  consistent and invertible estimates of  $M_1$  and  $M_2$ , respectively, and with  $\approx$  meaning asymptotic equivalence:  $X \approx Y$  iff  $X = Y(1 + o_p(1))$ . In this case,  $\hat{\theta}$  is an (asymptotically) efficient pooled estimator (as is well known and can be shown via Stuart, Ord and Arnold 1999, p. 103, for example) of  $\theta$ , and for tests  $U$  and  $W$  let:

$$U \approx \hat{\theta}' \hat{V}_{\hat{\theta}}^{-1} \hat{\theta}, \quad W \approx (\hat{\theta}_1 - \hat{\theta}_2)' \hat{V}_{\hat{\theta}_1 - \hat{\theta}_2}^{-1} (\hat{\theta}_1 - \hat{\theta}_2),$$

with variance-covariance estimators  $\hat{V}_{\hat{\theta}} \approx n^{-1}(\hat{M}_1^{-1} + \hat{M}_2^{-1})^{-1}$  and  $\hat{V}_{\hat{\theta}_1 - \hat{\theta}_2} \approx n^{-1}(\hat{M}_1 + \hat{M}_2)$ . Note that, under  $H_0$ ,  $nE \hat{\theta}(\hat{\theta}_1 - \hat{\theta}_2)' \rightarrow 0_{p,p}$ , where  $0_{p,p}$  is the  $p \times p$  matrix consisting of 0's. Setting  $u \approx \hat{V}_{\hat{\theta}}^{-1/2} \hat{\theta}$  and  $w \approx \hat{V}_{\hat{\theta}_1 - \hat{\theta}_2}^{-1/2} (\hat{\theta}_1 - \hat{\theta}_2)$ , where  $\hat{V}^{-1/2} = (\hat{V}^{1/2})^{-1}$  and  $\hat{V}^{1/2}$  is the Cholesky root, Assumption 2 follows.

To obtain an analytic expression for asymptotic distributions we can further describe  $G$  as a quadratic form:

$$G \approx (u', w') A (u', w)'$$

with:

$$A = \begin{bmatrix} I & -\frac{a^*}{2} I \\ -\frac{a^*}{2} I & -b^* I \end{bmatrix},$$

where  $a^*$  and  $b^*$  are the large- $n$  limits of  $a$  and  $b$ , and  $I$  is the  $p \times p$  identity matrix. Since  $(u', w)'$  is (asymptotically) standard normal we can express  $G$  as a weighted sum of independent chi square variables (as in Scheffe 1959 and Imhof 1961), as follows:

$$G \xrightarrow{d} \sum_{r=1}^q \lambda_r \chi_{h_r, r}^2 \tag{4}$$

where  $\lambda_1, \dots, \lambda_q$  are the distinct eigenvalues (arranged in decreasing order) of  $A$ ,  $h_r$  is the multiplicity of the  $r$ -th eigenvalue, and the variables  $\chi_{h_r, r}^2$

are mutually independent chi square variables, having respective degrees of freedom  $h_r$ . For each eigenvalue  $\lambda$  the eigenvectors turn out to be of the form:  $x^{(1)} = (1, 0_{1,p-1}, c, 0_{1,p-1})'$ ,  $x^{(2)} = (0, 1, 0_{1,p-1}, c, 0_{1,p-2})'$ ,  $\dots$ ,  $x^{(p)} = (0_{1,p-1}, 1, 0_{1,p-1}, c)'$ , for some constant  $c$  and  $0_{1,p-1}$  the  $1 \times (p-1)$  vector consisting of 0's, etc., and setting  $Ax^{(k)} = \lambda x^{(k)}$  yields the equations  $1 - (a^*/2)c = \lambda$  and  $-a^*/2 - bc = \lambda c$ , from which we obtain:

$$\lambda = \frac{1}{2} \left( 1 - b^* \pm \sqrt{(1 + b^*)^2 + (a^*)^2} \right). \quad (5)$$

For each  $\lambda$  the associated  $c = (2/a^*)(1 - \lambda)$ , and  $q = 2$ ,  $h_1 = h_2 = p$ . Since  $b^* > 0$  we obtain  $\lambda_1 \geq 1 > 0$  and  $\lambda_2 < 0$ , hence the asymptotic distribution of  $G$  has full support  $(-\infty, \infty)$ .

For the nested-sample test,  $\lambda_1 = 1$  and  $\lambda_2 = -\frac{1-\rho}{\rho}$ . Consequently, the asymptotic (cumulative) distribution  $F_{\text{nest}}(\xi) = P(G_{\text{nest}} \leq \xi)$  is decreasing in  $\rho$  for each  $\xi$ . As  $\rho \rightarrow 1$ ,  $F$  approaches the  $\chi_p^2$  distribution.

For the split-sample test, eigenvalues are:

$$\lambda = \frac{1}{2} \left( -\frac{1}{\rho} \pm \sqrt{\frac{1}{\rho^2} + \frac{4}{\rho(1-\rho)}} \right).$$

As a function of  $\rho$ , the larger eigenvalue ( $\lambda_1$ ) is increasing, with  $\lambda_1 \downarrow 2$  as  $\rho \downarrow 0$  and  $\lambda_1 \uparrow \infty$  as  $\rho \uparrow 1$ ;  $\lambda_2$  is increasing for  $\rho < 2/3$ , equals  $-3$  at  $\rho = 2/3$ , and is decreasing for  $\rho > 2/3$ , with  $\lambda_2 \rightarrow -\infty$  as  $\rho \rightarrow 0$  or  $\rho \rightarrow 1$ . Consequently, the distribution  $F_{\text{split}}(\xi)$  of  $G_{\text{split}}$  is decreasing in  $\rho$  for  $\rho < 2/3$ ; however, for  $\rho$  approaching 1,  $G_{\text{split}}$  is approximately  $(\chi_{p,1}^2 - \chi_{p,2}^2)/\sqrt{1-\rho}$ , causing  $F_{\text{split}}(\xi)$  to increase in  $\rho$  at  $\xi < 0$ .

We can use (4) to express the expected value of test  $G$  as  $(\lambda_1 + \lambda_2)p$ , and for the nested-sample test this is  $\mu_{\text{nest}} = \frac{p(2\rho-1)}{\rho}$ , while for the split-sample test it is  $\mu_{\text{split}} = -\frac{p}{\rho}$ . For the significance test  $U$ ,  $EU = p > \mu_{\text{nest}} > \mu_{\text{split}}$ . The variance of  $G$  is  $2p(\lambda_1^2 + \lambda_2^2)$  (compared to  $V(U) = 2p$ ), and for  $G_{\text{nest}}$  this is  $\nu_{\text{nest}} = \frac{2p}{\rho^2}(\rho^2 + (1-\rho)^2)$ , while for  $G_{\text{split}}$  it is  $\nu_{\text{split}} = \frac{2p}{\rho^2} \frac{1+\rho}{1-\rho}$ . Consequently,

$V(U) < \nu_{\text{nest}} < \nu_{\text{split}}$ . For large  $p$  the distributions of  $U$ ,  $G_{\text{nest}}$  and  $G_{\text{split}}$  are approximately normal  $N(p, 2p)$ ,  $N(\mu_{\text{nest}}, \nu_{\text{nest}})$ ,  $N(\mu_{\text{split}}, \nu_{\text{split}})$ , respectively.

For the (asymptotic) correlations between the proposed tests  $G$  and the tests  $U$  and  $W$ , using (1) and Assumptions 1 and 2 we obtain  $\text{corr}(G, U) = (\lambda_1^2 + \lambda_2^2)^{-1/2} > 0$  and  $\text{corr}(G, W) = -b(\lambda_1^2 + \lambda_2^2)^{-1/2} < 0$ , under  $H_0$ . Hence, for the nested-sample test, as  $\rho \rightarrow 1$  the correlations with  $U$  and  $W$  approach 1 and 0, respectively, and as  $\rho \rightarrow 0$  the correlations approach 0 and -1. For the split-sample test:  $\text{corr}(G_{\text{split}}, U)$  is maximized at  $\rho = \frac{1}{2}(\sqrt{5} - 1) = 0.6180$  (to 4 decimals), and approaches 0 as  $\rho$  approaches 0 or 1;  $\text{corr}(G_{\text{split}}, W)$  approaches 0 as  $\rho \rightarrow 1$  and approaches  $-1$  as  $\rho \rightarrow 0$ .

To compute test distributions we use Imhof (1961, Section 3) to obtain:

$$P(G \leq \xi) = \frac{1}{2} - \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sin\left(\frac{\rho}{2}(\arctan(\lambda_1 u) + \arctan(\lambda_2 u)) - \frac{\xi u}{2}\right)}{u((1 + \lambda_1^2 u^2)(1 + \lambda_2^2 u^2))^{p/4}} du. \quad (6)$$

To this we apply numerical integration (Mathematica 4.0 NIntegrate tool). Table 1 reports critical values, to 4 decimal places, for  $\rho = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{6}{7}$ ,  $p = 1, \dots, 10$ , and significance levels  $\alpha = 0.10, 0.05, 0.01$ . We include more values of  $\rho$  near 1 because for cross-validation the estimation sub-sample is often a large portion of the sample. For  $p$  even we check the results by comparing them to an alternative formula for  $P(G \leq \xi)$ , using Imhof (1961, Section 2) and the Mathematica 4.0 symbolic derivative tool  $D[\cdot]$ . Also, we check all results by simulating the distribution (4) via a normal random number generator. From our previous discussion the nested-sample test critical values must increase in  $\rho$ , as they do in Table 1. For the split-sample test the critical values must increase for  $\rho < 2/3$ , as they do in Table 1 even for  $\rho \geq 2/3$ .

### 3. POWER

To describe the asymptotic power of the proposed tests, let  $H_{\text{hom}}$  denote the hypothesis that  $\theta$ 's value is constant (“homogeneous”) across the population, and that  $\theta \neq 0$ . Also, let  $H_{\text{het}}$  denote the hypothesis that the true  $\theta$  value differs (is “heterogeneous”) across the two (exclusive, exhaustive) subpopulations under study. We will show that the proposed tests of  $H_0 : \theta = 0$  have different behaviors under asymptotic-local versions of the two alternatives  $H_{\text{hom}}$  than  $H_{\text{het}}$ . The nested-sample test  $G_{\text{nest}}$ , and its generalization  $G_{\text{nest}}^*$ , has greater local power under  $H_{\text{hom}}$  than  $H_{\text{het}}$ , whereas classical significance tests have the same power in the two cases. This property of  $G_{\text{nest}}^*$  lets it focus on finding global theories, e.g. non-zero  $\theta$  values that are constant across the population. By comparison, the split-sample test  $G_{\text{split}}^*$  has a similar property, but to a more limited degree.

To specify a (local/asymptotically-vanishing) version of  $H_{\text{hom}}$ , with the population homogeneous with respect to  $\theta$  but not necessarily with respect to other parameters, we have:

*Assumption 3:*  $(u', w)'$  converges in distribution to  $N((\delta', 0, \dots, 0)', I)$ , for some  $p$ -vector  $\delta \neq 0$ .

This is a natural modification, for asymptotic-local ( $H_{\text{hom}}$ ) alternatives, of Assumption 2. For example, let the local alternative to  $H_0$  be that  $\theta = \omega/\sqrt{n}$  for a  $p$ -vector  $\omega$  having some non-zero elements, in which case, in the SETUP we satisfy Assumption 3 with  $\delta = ((M_1^{-1} + M_2^{-1})^{1/2})' \omega$ .

*Theorem 1:* Let Assumptions 1 through 3 hold, and for Assumption 3 let  $\delta = \gamma \delta^*$  for some constant  $\gamma \neq 0$  and some  $p$ -vector  $\delta^* \neq 0$ . Then each of the following is true of asymptotic test power:

- (i) For  $G_{\text{nest}}^*$ , power increases in  $\gamma$  (hence is unbiased) and also in  $\rho$  (approaching that of test  $U$ ).

- (ii)  $G_{\text{split}}^*$  is unbiased for  $\gamma$  sufficiently large, and at each  $\rho$  power is decreasing in  $\rho$  when  $\gamma$  is sufficiently large.
- (iii) For  $\gamma$  sufficiently large,  $G_{\text{nest}}^*$  has greater power than  $G_{\text{split}}^*$ .

The fact that the asymptotic-local power of  $G_{\text{nest}}^*$  improves as  $\rho \rightarrow 1$  seems intuitive because, with  $G_{\text{nest}}^* \approx U - ((1-\rho)/\rho)W$ , the “penalty” term  $-((1-\rho)/\rho)W$  (which is independent of  $U$  and invariant to  $\theta$  under Assumption 3) diminishes in importance; however, Theorem 1 relies on the (asymptotic) chi square distributions of  $U$  and  $W$ . If, say, for scalar  $\theta$  and hypothesis  $\theta = 0$  a(n upper-tailed test) statistic  $x$  has density  $f(x) = -(x-\theta)$  for  $x \in [\theta-1, \theta]$ ,  $f(x) = x - \theta$  for  $x \in [\theta, \theta + 1]$ ,  $f(x) = 0$  otherwise, and  $y$  has distribution  $P(y = -1/2) = P(y = 1/2) = 1/2$ , independent of  $x$ , then for the sum  $x + y$ , at test size  $\alpha = 0.5$  (hence critical value = 0) the upper-tailed  $x + y$  test of  $\theta = 0$  has greater power (=3/4) when  $\theta = 0.5$  than does the upper-tailed  $x$  test (power = 5/8), despite the fact that  $y$  is “noise” added to  $x$ .

The greater power of the nested-sample test, relative to the split-sample test, can be intuitively understood from an equivalence of  $G_{\text{nest}}^*$  to nested-sample cross-validation (Section 4) which uses more information (for validation) than does the split-sample scheme. Both methods use less information than a full-sample scheme, consistent with the fact that the nested- and split-sample tests have less power than the full-sample test  $U$ . We can illustrate bias in  $G_{\text{split}}^*$ , with  $p = 1$ ,  $\rho = 0.99$ ,  $\delta = 5$ ,  $\alpha = 0.9$ . Asymptotically, under Assumptions 1 and 3,  $G_{\text{split}}^*$  is distributed as  $(z_1 + \delta)^2 - a(z_1 + \delta)z_2 - bz_2^2$ , with  $z_1, z_2$  independent standard normal variables, and we compute (via simulation) the rejection rate = 0.68 <  $\alpha$ . In the same setting, the rejection rate for  $G_{\text{nest}}^*$  is 1.00 (to 2 decimal places).

To accommodate alternatives exhibiting population heterogeneity, we have:

*Assumption 4:*  $(u', w')'$  converges in distribution to  $N(\mu, I)$  with  $\mu = (\mu'_1, \mu'_2)'$

and  $p$ -vectors  $\mu_1, \mu_2$  such that  $\mu_2$  has some non-zero elements.

This is suited to asymptotic-local heterogeneous ( $H_{\text{het}}$ ) alternatives. To illustrate consider the SETUP, with  $\theta = \omega_k/\sqrt{n}$  on the  $k$ -th sub-sample,  $k = 1, 2$ . Then  $\mu_1 = ((M_1^{-1} + M_2^{-1})^{1/2})' (M_1^{-1} + M_2^{-1})^{-1} (M_1^{-1}\omega_1 + M_2^{-1}\omega_2)$  and  $\mu_2 = (M_1 + M_2)^{-1/2}(\omega_1 - \omega_2)$ .

*Theorem 2:* Let Assumptions 1 and 2 hold, and in the specification of Assumption 3 let  $\delta = \mu_1$ . Then each of the following holds for asymptotic test power:

- (i)  $G_{\text{nest}}^*$  has a lower rejection probability under intra-population parameter differences (Assumption 4) than under parameter constancy (Assumption 3).
- (ii) For all  $\rho$  sufficiently small,  $G_{\text{split}}^*$  has a lower rejection probability under parameter differences than under parameter constancy.
- (iii) The full-sample test  $U$  has the same power under parameter differences and parameter constancy.

This theorem formalizes a sense in which  $G^*$  test power is lower in the presence of parameter change, allowing the test to discriminate between homogeneous and heterogeneous alternatives. By comparison, the test  $U$  is invariant to such change because the relevant signal depends on a (weighted) average of  $\theta$  values across regimes, but not on regime differences. While the power drop effect applies broadly to the nested-sample test, for the split-sample test it is guaranteed only for small  $\rho$ ; for  $\rho$  near 1 parameter change can drop or raise  $G_{\text{split}}^*$ 's power. If, say  $p = 1$ ,  $\alpha = 0.1$ ,  $\rho = 0.9$  and  $\mu_1 = 5$  then local power at values  $\mu_2 = -1, 0, 1$  is (computed via Table 1 and simulation of  $(z_1 + \mu_1)^2 - a(z_1 + \mu_1)(z_2 + \mu_2) - b(z_2 + \mu_2)^2$ , with  $z_1, z_2$  independent  $N(0,1)$ ) 0.99, 0.95, 0.77, respectively.

#### 4. CROSS-VALIDATION

To link the  $G^*$  tests to the idea of cross-validated model performance, for a data sequence of random vectors  $z_1, \dots, z_n$  consider the maximization of a generalized log-likelihood function  $\mathcal{L}(\psi) = \sum_{i=1}^n g(z_i; \psi)$ , for some function  $g$  and  $r$ -vector  $\psi$  (with  $r \geq p$ ) which we can write as  $\psi = (\theta', \nu')$ , for some  $(r - p) \times 1$  vector  $\nu$ . If  $z_1, \dots, z_n$  are independent and identically distributed (iid) then we can set  $g = \ln(f)$ , with  $f$  a probability mass or density function for  $z$ . For a stationary Markov sequence  $y_0, y_1, \dots, y_n$  we can let  $z_i = (y_i, y_{i-1})$ ,  $i = 1, \dots, n$ , and let  $g = \ln(f)$  with  $f$  the conditional density or probability mass of  $y_i$  given  $y_{i-1}$ . Defining sub-sample index sets  $S_1: \{i = 1, \dots, m\}$  and  $S_2: \{i = m + 1, \dots, n\}$  we have generalized log-likelihoods  $\mathcal{L}_k(\psi_k) = \sum_{i \in S_k} g(z_i, \psi_k)$ . The true value of sub-vectors  $\theta_k$  of  $\psi_k$  can differ across  $k$ , but for the desired link to cross-validation we suppose that the  $\nu_k$  true values are the same across  $k$ .

Suppose that there exists both a unique unconstrained  $\mathcal{L}$  maximizer  $\hat{\psi}$  and a constrained maximizer  $\tilde{\psi}$ , with constraint  $\theta = 0$ . Let  $\hat{\psi}_k$  and  $\tilde{\psi}_k$ ,  $k = 1, 2$ , be the corresponding sub-sample estimators on  $S_k$ . To validate the restriction  $H_0: \theta = 0$  consider the statistics:

$$G_{\text{nest}}^\dagger = 2 \left( \mathcal{L}(\hat{\psi}_1) - \mathcal{L}(\tilde{\psi}_1) \right), \quad G_{\text{split}}^\dagger = 2 \left( 1 - \frac{m}{n} \right)^{-1} \left( \mathcal{L}_2(\hat{\psi}_1) - \mathcal{L}_2(\tilde{\psi}_1) \right), \quad (7)$$

where  $G_{\text{nest}}^\dagger$  uses a nested scheme of training and validation samples to assess  $H_0$ , and  $G_{\text{split}}^\dagger$  uses a split-sample scheme. For  $G_{\text{nest}}^\dagger$  the formula in (7) reduces to the full-sample likelihood ratio test statistic if we set  $m = n$ ; for  $G_{\text{split}}^\dagger$  the factor  $(1 - m/n)^{-1}$  reflects the fact that validation here uses only  $100(1 - m/n)$  percent of the data.

To proceed, let the asymptotic covariance matrices  $M_1$  and  $M_2$  of sub-sample estimators  $\hat{\theta}_k, k = 1, 2$  be equal, up to (sub-)sample size effects, as follows:

$$\rho M_1 = M = (1 - \rho)M_2, \quad (8)$$

for some invertible  $M$ . Also, let:

$$\hat{\theta}_1 \approx \hat{\theta} + (1 - \rho) (\hat{\theta}_1 - \hat{\theta}_2), \quad \hat{\theta}_2 \approx \hat{\theta} - \rho (\hat{\theta}_1 - \hat{\theta}_2), \quad (9)$$

which holds provided that  $\hat{\theta} \approx \rho \hat{\theta}_1 + (1 - \rho) \hat{\theta}_2$ , this being widely applicable under (8). Next, under  $H_0$  or a local alternative (Assumption 3 or 4), let:

$$\mathcal{L}_k(\hat{\psi}_1) - \mathcal{L}_k(\psi^*) \approx a'_k(\hat{\psi}_1 - \psi^*) - \frac{1}{2}(\hat{\psi}_1 - \psi^*)' F_k(\hat{\psi}_1 - \psi^*), \quad k = 1, 2, \quad (10)$$

with  $\psi^* = (0, \nu)'$ ,  $a_k = \sum_{i \in S_k} \frac{\partial}{\partial \psi} g(z_i; \psi^*)$ , and  $F_k = - \sum_{i \in S_k} \frac{\partial^2}{\partial \psi \partial \psi'} g(z_i; \psi^*)$ . Also, if  $r > p$  let:

$$\mathcal{L}_k(\tilde{\psi}_1) - \mathcal{L}_k(\psi^*) \approx a'_{k\nu}(\tilde{\nu}_1 - \nu) - \frac{1}{2}(\tilde{\nu}_1 - \nu)' F_{k\nu\nu}(\tilde{\nu}_1 - \nu), \quad k = 1, 2, \quad (11)$$

with  $a_{k\nu}$  the lower  $(r - p) \times 1$  sub-vector of  $a_k$ , and  $F_{k\nu\nu}$  the lower-right  $(r - p) \times (r - p)$  sub-matrix of  $F_k$ . Further, with  $n_1 = m$  and  $n_2 = n - m$  the two sub-sample sizes (for sub-samples  $i = 1, 2, \dots, m$  and  $i = m + 1, m + 2, \dots, n$ ), let:

$$\frac{F_k}{n_k} \xrightarrow{p} F, \quad \hat{\psi}_k - \psi^* \approx n_k^{-1} F^{-1} a_k, \quad n_k^{-1/2} a_k \xrightarrow{d} N(F(\omega'_k, 0_{1,q})', F), \quad (12)$$

and if  $r > p$  let:

$$\tilde{\nu}_k - \nu \approx n_k^{-1} F_{\nu\nu}^{-1} a_{k\nu}, \quad (13)$$

where  $F$  is a positive definite matrix with lower-right  $(r - p) \times (r - p)$  sub-matrix  $F_{\nu\nu}$ , and  $\omega_1, \omega_2$  are regime-specific ‘local’ effect  $p \times 1$  vectors.

*Assumption 5:* Let (8, 9, 10, 12) hold, and if  $r > p$  let (11) and (13) hold. Moreover, let  $a_1$  and  $a_2$  be asymptotically independent random vectors.

This allows a variety of data designs and models (cross-section, time series, etc.), and the technical conditions on the likelihood and its derivatives are standard (see for example Schervish 1995, Ch. 7.3.5). To specify the  $G^*$  tests to which we will compare  $G^\dagger$ , we start with our quadratic forms  $G_{\text{nest}}$  and  $G_{\text{split}}$ , such that  $u = \hat{V}_{\hat{\theta}}^{-1/2}\hat{\theta}$  and  $w = \hat{V}_{\hat{\theta}_1 - \hat{\theta}_2}^{-1/2}(\hat{\theta}_1 - \hat{\theta}_2)$ , as described in the SETUP (Section 2), and set  $\hat{V}_{\hat{\theta}} = n^{-1}(\hat{M}_1^{-1} + \hat{M}_2^{-1})^{-1}$  and  $\hat{V}_{\hat{\theta}_1 - \hat{\theta}_2} = n^{-1}(\hat{M}_1 + \hat{M}_2)$ , where  $\hat{M}_1$  and  $\hat{M}_2$  are any consistent estimators of  $M$  obtained from the two sub-samples (such as those obtained from likelihood Hessians  $F_1$  and  $F_2$ ).

*Theorem 3:* Under Assumptions 1, 2, 5 and either  $H_0$  or its alternatives (Assumption 3 or 4),  $G_{\text{nest}}^* \approx G_{\text{nest}}^\dagger$  and  $G_{\text{split}}^* \approx G_{\text{split}}^\dagger$ .

Theorem 3 relies on large-sample asymptotics but in some cases there is an exact finite-sample relationship between the quadratic form  $G$  (of  $G^*$ ) and the “cross-validating” statistic  $G^\dagger$ , as in:

*Example 1:*  $z_1, \dots, z_n$  are mutually independent and  $N(\theta_k, 1)$  in the  $k$ -th sub-population. With  $g(z; \theta) = -\ln(2\pi)/2 - (z - \theta)^2/2$ , unconstrained maximum likelihood estimators (mle’s)  $\hat{\theta}_k$ ,  $k = 1, 2$  are the sub-sample averages  $\bar{z}_1$  and  $\bar{z}_2$  on  $S_1$  and  $S_2$ , respectively, constrained mle’s = 0, and  $M = 1 = p = r = F$ ,  $G_{\text{nest}}^\dagger = \sum_{i=1}^n z_i^2 - (z_i - \bar{z}_1)^2$  and  $G_{\text{split}}^\dagger = (1 - m/n)^{-1} \sum_{i=m+1}^n z_i^2 - (z_i - \bar{z}_1)^2$ . Expanding squares and simplifying,  $G_{\text{nest}}^\dagger = n(2\hat{\theta}_1\hat{\theta} - \hat{\theta}_1^2)$  and  $G_{\text{split}}^\dagger = n(2\hat{\theta}_1\hat{\theta}_2 - \hat{\theta}_1^2)$ . Using  $\hat{\theta}_1 = \hat{\theta} + (1 - \frac{m}{n})(\hat{\theta}_1 - \hat{\theta}_2)$  and  $\hat{\theta}_2 = \hat{\theta} - \frac{m}{n}(\hat{\theta}_1 - \hat{\theta}_2)$ , and defining  $u = \sqrt{n}\hat{\theta}$ ,  $w = (1/m + 1/(n - m))^{-1/2}(\hat{\theta}_1 - \hat{\theta}_2)$ , we obtain  $G_{\text{nest}} = n\hat{\theta}^2 - ((n - m)^2/n)(\hat{\theta}_1 - \hat{\theta}_2)^2 = G_{\text{nest}}^\dagger$  and  $G_{\text{split}} = n\hat{\theta}^2 - 2m\hat{\theta}(\hat{\theta}_1 - \hat{\theta}_2) - (n - m^2/n)(\hat{\theta}_1 - \hat{\theta}_2)^2 = G_{\text{split}}^\dagger$ .

For the Bernoulli, exponential and Poisson (one-parameter) distributions, asymptotic (but not finite-sample) equivalences in Theorem 3 can be obtained directly (by evaluating  $G$ ’s and  $G^\dagger$ ’s, details omitted for brevity). For a two-parameter distribution consider:

*Example 2:*  $z_1, \dots, z_n$  are mutually independent and  $N(\theta_k, \sigma^2)$  on the  $k$ -th sub-population. With  $g(z; \theta) = -(1/2)(\ln(2\pi) + \ln(\sigma^2) + (z - \theta)^2/\sigma^2)$ , mle's of  $\theta$  are as in Example 1,  $r = 2$ ,  $p = 1$ ,  $M = \sigma^2 = \nu$ ,  $\tilde{\nu}_1 = m^{-1} \sum_1^m z_i^2$ ,  $\hat{\nu}_1 = m^{-1} \sum_1^m (z_i - \bar{z}_1)^2$ , Hessian matrix  $F$  is well-known (see Stuart, Ord and Arnold 1999, p. 75) and:

$$G_{\text{nest}}^\dagger = n \ln \left( \frac{\sum_1^m z_i^2}{\sum_1^m (z_i - \bar{z}_1)^2} \right) + \frac{\sum_1^n z_i^2}{m^{-1} \sum_1^m z_i^2} - \frac{\sum_1^n (z_i - \bar{z}_1)^2}{m^{-1} \sum_1^m (z_i - \bar{z}_1)^2},$$

$$G_{\text{split}}^\dagger = n \ln \left( \frac{\sum_1^m z_i^2}{\sum_1^m (z_i - \bar{z}_1)^2} \right) + \left(1 - \frac{m}{n}\right)^{-1} \left( \frac{\sum_{m+1}^n z_i^2}{m^{-1} \sum_1^m z_i^2} - \frac{\sum_{m+1}^n (z_i - \bar{z}_1)^2}{m^{-1} \sum_1^m (z_i - \bar{z}_1)^2} \right).$$

We have  $\ln\left(\frac{\sum_{i=1}^m z_i^2}{\sum_{i=1}^m (z_i - \bar{z}_1)^2}\right) \approx (\bar{z}_1)^2/\sigma^2$  under  $H_0$  and local alternatives, and for the last two quotients in each  $G^\dagger$  expression we can write their difference as  $c/d - e/f = (1/d)(c - e) + e(1/d - 1/f)$ : For  $G_{\text{nest}}^\dagger$ ,  $(1/d)(c - e) \approx n(2\bar{z}_1\bar{z} - \bar{z}_1^2)/\sigma^2$ ,  $e(1/d - 1/f) \approx -n(\bar{z}_1)^2/\sigma^2$ , in which case  $G_{\text{nest}}^\dagger \approx n(2\hat{\theta}_1\hat{\theta} - \hat{\theta}_1^2)/\sigma^2$ . Setting  $u = \sqrt{n/\hat{\nu}}\hat{\theta}$  and  $w = ((\hat{\nu}(1/m+1/(n-m)))^{-1/2}(\hat{\theta}_1 - \hat{\theta}_2))$ , and applying (9) we obtain  $G_{\text{nest}} \approx G_{\text{nest}}^\dagger$ . Similarly,  $G_{\text{split}} \approx G_{\text{split}}^\dagger$ .

## 5. REGRESSION

As Example 2 illustrates, when testing regression coefficients the proposed test statistics  $G_{\text{nest}}$  and  $G_{\text{split}}$  typically differ in finite samples from the ‘‘cross-validating’’ test statistics  $G_{\text{nest}}^\dagger$  and  $G_{\text{split}}^\dagger$ , even under the simplifying assumptions (Assumption 5) that deliver large-sample equivalence between the two types of test ( $G$  and  $G^\dagger$ ). To more fully illustrate the exact behavior of the  $G$  tests, consider the linear regression model:

$$y_i = \theta'x_i + \gamma'v_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (14)$$

with non-stochastic  $p$ -vectors  $x_i$  and  $q$ -vectors  $v_i$ , and errors  $\varepsilon_i$  iid  $N(0, \sigma^2)$ . Then  $\nu = (\gamma', \sigma^2)'$ , and with  $Z$  the  $n \times (p + q)$  matrix with  $i$ -th row  $Z_i =$

$(x'_i, v'_i)$ , let  $m^{-1} \sum_{i=1}^m Z'_i Z_i = (n-m)^{-1} \sum_{i=m+1}^n Z'_i Z_i = L$  for an invertible matrix  $L$ . Let  $\hat{\theta}$  be the full-sample ordinary least squares (OLS) estimator, let  $\hat{V}_{\hat{\theta}}$  be the OLS covariance matrix estimator for  $\hat{\theta}$ , and let  $\hat{V}_{\hat{\theta}_1 - \hat{\theta}_2} = (n/m + n/(n-m))\hat{V}_{\hat{\theta}}$ . With this regression setup we can readily apply Assumptions 2 through 4 and hence Theorems 1 and 2 regarding the statistics  $G$ , and setting  $g(z_i, \psi) = -(1/2)(\ln(2\pi) + \ln(\sigma^2) + (y_i - \theta'x_i - \gamma'v_i)^2/\sigma^2)$  we can verify Assumption 5 and hence apply Theorem 3. Further, we have  $\hat{\theta}_1 = \hat{\theta} + (1 - m/n)(\hat{\theta}_1 - \hat{\theta}_2)$  and  $\hat{\theta}_2 = \hat{\theta} - (m/n)(\hat{\theta}_1 - \hat{\theta}_2)$ , and with specification  $u = \hat{V}_{\hat{\theta}}^{-1/2}\hat{\theta}$  and  $w = \hat{V}_{\hat{\theta}_1 - \hat{\theta}_2}^{-1/2}(\hat{\theta}_1 - \hat{\theta}_2)$  we obtain:

$$G_{\text{nest}} = \hat{\sigma}^{-2} \sum_1^n (y_i - \tilde{\gamma}'_1 v_i)^2 - (y_i - \hat{\theta}'_1 x_i - \hat{\gamma}'_1 v_i)^2, \quad (15)$$

$$G_{\text{split}} = (1 - m/n)^{-1} \hat{\sigma}^{-2} \sum_{m+1}^n (y_i - \tilde{\gamma}'_1 v_i)^2 - (y_i - \hat{\theta}'_1 x_i - \hat{\gamma}'_1 v_i)^2, \quad (16)$$

with  $\hat{\sigma}$  the OLS regression standard error. To show (15) and (16) one can begin in the easy case of orthonormal regressor columns ( $L =$  identity matrix), then verify (via straightforward algebra) invariance with respect to transformations  $Z \rightarrow ZJ$  for invertible  $(p+q) \times (p+q)$  matrices  $J$ .

Gilbert (2001) first proposed the “nested-sample F test” statistic given by formula (15) (upon division by  $p$ ). McCracken (1999) and Gilbert (2001) proposed the “split-sample F test” given by (16) (again, upon division by  $p$ ). Clark and McCracken (2001a) study some related tests (and see Good 2001, Ch. 10 for some other discussion). By comparison, the  $F$  statistic for testing  $H_0$  is  $F = p^{-1} \hat{\sigma}^{-2} \sum_1^n ((y_i - \tilde{\gamma}'_1 v_i)^2 - (y_i - \hat{\theta}'_1 x_i - \hat{\gamma}'_1 v_i)^2)$  and (excepting division by  $p$ ) differs from  $G_{\text{nest}}$  only in use of the full sample rather than a sub-sample for estimation, but differs from  $G_{\text{split}}$  both in estimation and validation sample choices.

The finite-sample distribution of the  $G$  regression statistics, under  $H_0$ , is as follows:

*Theorem 4:* Under  $H_0$  in the regression model (14), for  $p \times 1$  vectors  $\zeta_1$  and  $\zeta_2$ , and  $(n - 2K - L) \times 1$  vector  $\zeta_3$ , such the elements of  $\zeta_1, \zeta_2, \zeta_3$  are all mutually independent  $N(0,1)$  variables, each of the following holds:

$$G_{\text{nest}} \stackrel{d}{=} (n - p - q) \times \frac{2\sqrt{\frac{n-m}{m}} \sum_{k=1}^p Z_{1k}Z_{2k} + \left(1 - \frac{n-m}{m}\right) \sum_{k=1}^p Z_{1k}^2}{\sum_{k=1}^p \left(\sqrt{\frac{n-m}{n}} Z_{1k} - \sqrt{\frac{m}{n}} Z_{2k}\right)^2 + \sum_{k=1}^{n-2p-q} Z_{3k}^2}, \quad (17)$$

$$G_{\text{split}} \stackrel{d}{=} \frac{n - p - q}{1 - \frac{m}{n}} \times \frac{2\sqrt{\frac{n-m}{m}} \sum_{k=1}^p Z_{1k}Z_{2k} - \frac{n-m}{m} \sum_{k=1}^p Z_{1k}^2}{\sum_{k=1}^p \left(\sqrt{\frac{n-m}{n}} Z_{1k} - \sqrt{\frac{m}{n}} Z_{2k}\right)^2 + \sum_{k=1}^{n-2p-q} Z_{3k}^2}. \quad (18)$$

To interpret these distributions we can compare them to the  $F$  distribution, the latter being that of the variable:

$$\frac{n - p - q}{p} \frac{\sum_{i=1}^p Z_{4k}^2}{\sum_{i=1}^{n-p-q} Z_{5k}^2},$$

where the elements of  $Z_4$  and  $Z_5$  are all iid  $N(0, 1)$ . The elements of  $Z_4$  are interpretable as standardized deviations of (full-sample) parameter estimates from their true values, while the elements of  $Z_5$  are interpretable as variables comprising the remaining degrees of freedom in the data. By comparison, in (17) and (18),  $Z_1$  and  $Z_2$  are interpretable as standardized deviations of parameter estimates from their true values (see Appendix), with  $Z_1$  obtained from OLS estimation on  $(v_i, x_i, y_i, i = 1, \dots, m)$ , and  $Z_2$  obtained from  $(v_i, x_i, y_i, i = m + 1, \dots, n)$ . The vector  $Z_3$  consists of variables comprising the remaining degrees of freedom in the data, and so the distributions (17) and (18) are similar to that of the  $F$  test, but somewhat more complex. For the formula (17), if we set  $m = n$  then  $G_{\text{nest}} \stackrel{d}{=} pF$  (and from (15),  $G_{\text{nest}} = pF$ ), but in that case the interpretation of  $Z_2$  as a sub-sample estimator breaks down.

In large samples, Theorem 4 yields:

$$G_{\text{nest}} \stackrel{d}{\approx} 2\sqrt{\frac{1-\rho}{\rho}} \sum_{k=1}^p Z_{1k}Z_{2k} + \left(1 - \frac{1-\rho}{\rho}\right) \sum_{k=1}^p Z_{1k}^2, \quad (19)$$

and:

$$G_{\text{split}} \stackrel{d}{\approx} (1-\rho)^{-1} \left( 2\sqrt{\frac{1-\rho}{\rho}} \sum_{k=1}^p Z_{1k}Z_{2k} - \frac{1-\rho}{\rho} \sum_{k=1}^p Z_{1k}^2 \right), \quad (20)$$

under  $H_0$ . These large-sample distributions are the same as those which obtain from the results of Section 2, as can be seen by making the substitution  $u \approx \rho^{1/2}Z_1 + (1-\rho)^{1/2}Z_2$ ,  $w \approx (\rho^{-1/2}Z_1 - (1-\rho)^{-1/2}Z_2)/\sqrt{\rho^{-1} + (1-\rho)^{-1}}$ .

Table 2 reports rejection rates for the  $G$  tests, using critical values from Table 1, and for a full-sample test  $U$  (= F test), using 10,000 simulation rounds. Here  $p = 1$ ,  $q = 2$ ,  $x_i$  and  $v_{i2}$  mutually independent standard normal sequences, and  $v_{i1} = \dots = v_{n1} = 1$ , for  $n = 100$  and  $n = 200$ . We report rejection rates under three hypotheses:  $H_0: \theta = 0$ ,  $H_{\text{hom}}: \theta = \frac{1}{4}$  and  $H_{\text{het}}: \theta$  equals 0 on the first sub-sample and equals  $(4(1 - m/n))^{-1}$  on the second sub-sample, with  $m/n = 1/4, 1/2, 3/4$ . With  $G$  statistics asymptotically equivalent to the corresponding  $G^\dagger$  statistics (Theorem 3), we do find  $G^\dagger$  tests to give similar results, omitted for brevity.

From Table 2, Under the null the  $F$  test rejects more frequently than the other tests (consistent with Theorem 1) except for the nested-sample test ( $m/n = 3/4$ ), which performs comparably. For each  $m/n$  the nested-sample test rejects more than the split-sample test does (consistent with Theorem 1), and overall the split-sample tests suffer considerable loss in power, relative to the full-sample and nested-sample tests. For the nested-sample test, under the null a higher  $m/n$  yields more frequent rejection, while for the split-sample test there is more frequent rejection at  $m/n = \frac{1}{4}, \frac{1}{2}$  than at  $\frac{3}{4}$  (also consistent with Theorem 1). Under parameter change, rejection rates of the proposed tests are lower, compared to results under the null (consistent with Theorem 2). The  $F$  test also rejects less in the presence of parameter differences, but the effect diminishes in the larger sample (consistent with

$F$  having the same local power under both alternatives). At higher values of  $m/n$  ( $= 4/5, \dots, 6/7$ ) the situation (omitted, for brevity) is qualitatively similar except that the split-sample test sometimes has more power in the presence of intra-population parameter differences than under population homogeneity, more so for higher  $m/n$ . This tendency at higher  $m/n$  is reversed if we switch the sign of  $\theta$  on the second sub-sample (consistent with the discussion following Theorem 2, and see Clark and McCracken 2001b for similar simulation results for other split-sample tests).

## 6. EXAMPLE

For an example with data consider the U.S. inflation rate  $y_i$ , given by the monthly percent change in consumer price index (all urban consumers), and the (civilian) unemployment rate  $z_i$ , each seasonally adjusted monthly series for the period February 1948 - January 2003 (data obtained from the FRED website, Federal Reserve Bank of St. Louis). A simple dynamic model of inflation is the regression:

$$y_i = \alpha + \beta x_{i-1} + \gamma y_{i-1} + \varepsilon_i, \quad i = 2, 3, \dots, n. \quad (21)$$

Let the parameter vector  $\theta$  of interest be the (scalar) coefficient  $\beta$ . Table 3 reports OLS estimates of the model for the two sub-periods 1948:03-1969:12 and 1970:01-2003:01, as well as various tests of  $H_0: \theta = 0$  and of  $\theta$  constancy over the two sub-periods. For our methods we choose the first sample period as the “training/estimation” sample (sample size  $= n_1 = 261$ ), and the latter sample period as the “validation” sample ( $n_2 = 391$ ). Under  $H_0$  the partial correlation (net of lagged  $y$ ) between inflation  $y_i$  and the past unemployment rate  $x_{i-1}$  is zero, whereas various economic theories suggest departures from  $H_0$  (see Romer 2001 for recent review and discussion). In the first sub-period

the  $\theta$  estimate is negative, consistent with the idea that low unemployment is associated with inflationary demand shocks (driving up demand in excess of the economy’s potential level of output), while in the second sub-period the  $\theta$  estimate is positive, consistent with stagflation in which negative supply shocks (including high oil costs) fuel inflation and reduce profits and jobs.

The proposed  $G$  tests, which are designed to reject more frequently when  $\theta \neq 0$  is constant across time than otherwise, show (in Table 3) no significant evidence against  $H_0$  (p-values  $\geq 0.97$ ), when using asymptotic critical values (Section 2). This is reasonable given the highly significant parameter change (reported via  $W$ ), and less significant  $\theta$  estimate (reported via  $U$ ). The  $G$  tests are obtained via  $u = \hat{\theta}/s_{\hat{\theta}}$  and  $w = (\hat{\theta}_1 - \hat{\theta}_2)/s_{\hat{\theta}_1 - \hat{\theta}_2}$  specified in two ways: (a) weighted least squares (WLS) approach (“weight”), where  $\hat{\theta}$  is the WLS estimator for  $\theta$  based on OLS  $\hat{\theta}_k$  and standard errors  $s_{\hat{\theta}_k}$ ,  $k = 1, 2$ ,  $s_{\hat{\theta}}$  is the WLS standard error, and  $s_{\hat{\theta}_1 - \hat{\theta}_2}^2 = s_{\hat{\theta}_1}^2 + s_{\hat{\theta}_2}^2$ ; (b) simple OLS approach (“simple”) with  $\hat{\theta}$  the full-sample OLS estimator,  $s_{\hat{\theta}}$  its standard error, and  $s_{\hat{\theta}_1 - \hat{\theta}_2}^2 = (n/m + n/(n - m))s_{\hat{\theta}}^2$ , in which case  $G$ ’s are given by (15) and (16). With the OLS approach, the split-sample  $G$  becomes the Gilbert-McCracken split-sample  $F$  statistic, and the nested-sample  $G$  becomes Gilbert’s (2001) nested-sample  $F$  statistic.

We also report  $G^\dagger$  statistics, obtained from split-sample or nested-sample likelihood evaluation (via Gaussian conditional density  $g(z_i, \psi)$  as specified in Section 5). For testing we use the asymptotic critical values (and  $G$ - $G^\dagger$  asymptotic equivalence, Theorem 3), and with this approach test results agree with the  $G$  statistics in finding no significant evidence of a stable non-zero value of  $\theta$  over time. We note however that our proof of asymptotic equivalence of  $G$  and  $G^\dagger$  relied on constancy of some nuisance parameters (second moments). It is easy to show non-equivalence when no such constancy is available. For our inflation model, regression standard errors differ notably across the two sub-samples, in which case the null (asymptotic) dis-

tribution of  $G^\dagger$  may differ substantially from that of  $G$ .

To summarize, the proposed methods do not find in the inflation model a way for the unemployment rate to explain future movements in inflation - consistently over the historical period. That is, a rather simplistic “Phillips curve” model like (21) does not appear to provide a global theory of inflation. Economists, who began noticing this instability of the Phillips curve in the 1970’s, have sought to build global theories (with supply- and demand-driven inflation sources, see Romer 2001). The proposed methods can likewise be applied to these more sophisticated models.

## 7. CONCLUSION

We compute test critical values based on (first-order) asymptotic theory, but in applications some second-order (Bartlett, etc.) corrections may be useful. Also, while we assume that underlying statistics ( $u$  and  $w$ ) conform to standard (normal) central limit theory, for some non-stationary data other limit distributions may apply, and critical values can be adjusted accordingly. Under simplifying assumptions the proposed tests are (asymptotically) equivalent to methods involving cross-validation, but as in the example in Section 6, the two sorts of statistics ( $G$  and  $G^\dagger$ ) can have large numerical differences. Such discrepancies can arise due to intra-population differences in nuisance parameters, and it would be interesting to study the issue in more detail.

## APPENDIX

Before proving Theorem 1 we will first establish two Lemmas. Let  $z_1$  and  $z_2$  be mutually independent standard normal  $p$ -vectors. For a  $p$ -vector  $\delta$  of the form  $\gamma\delta^*$  with scalar  $\gamma > 0$  and  $p$ -vector  $\delta^*$  having some non-zero elements define  $\pi(\gamma, b) = P((z_1 + \gamma\delta^*)'(z_1 + \gamma\delta^*) - bz_2'z_2 > c_{ab})$ , with  $b \geq 0$  and  $c_{ab}$  such that  $P(z_1'z_1 - bz_2'z_2 > c_{ab}) = \alpha$ . Let  $\pi_\gamma$  and  $\pi_b$  be the partial derivatives of  $\pi$ .

*Lemma 1:* Each of the following holds: (i)  $\pi_\gamma > 0$ , and (ii)  $\pi_b < 0$ .

Proof: Write  $\pi(\gamma, b) = \int (1 - F_{\chi^2(\lambda)}(c_{ab} + bv))f_{\chi^2}(v)dv$ , with  $F_{\chi^2(\lambda)}$  the non-central chi square cumulative distribution function with non-centrality parameter  $\lambda = \delta'\delta$ , and  $f_{\chi^2}$  the central chi square density, each with  $p$  degrees of freedom. With  $\pi(0, b) = \alpha$  and the fact that  $F_{\chi^2(\lambda)}(x)$  is decreasing in  $\lambda$  at each  $x$  (see Johnson and Kotz 1970, p. 135), (i) follows. For (ii) compute:

$$\pi_b = - \left( \frac{\partial}{\partial b} c_{ab} \right) \int f_{\chi^2(\lambda)}(c_{ab} + bv)f_{\chi^2}(v)dv - \int v f_{\chi^2(\lambda)}(c_{ab} + bv)f_{\chi^2}(v)dv,$$

with  $f_{\chi^2(\lambda)}$  the non-central chi square density. Differentiating (with respect to  $b$ ) on both sides of  $P(z_1'z_1 - bz_2'z_2 > c_{ab}) = \alpha$  we obtain:

$$\frac{\partial}{\partial b} c_{ab} = - \frac{\int v f_{\chi^2}(c_{ab} + bv)f_{\chi^2}(v)dv}{\int f_{\chi^2}(c_{ab} + bv)f_{\chi^2}(v)dv},$$

in which case:

$$\pi_b = \frac{\int f_{\chi^2(\lambda)}(c_{ab} + bv)f_{\chi^2}(v)dv}{\int f_{\chi^2}(c_{ab} + bv)f_{\chi^2}(v)dv} \int v f_{\chi^2}(c_{ab} + bv)f_{\chi^2}(v)dv - \int v f_{\chi^2(\lambda)}(c_{ab} + bv)f_{\chi^2}(v)dv,$$

so for  $\pi_b < 0$  it suffices that:

$$\frac{\int v f_{\chi^2}(c_{ab} + bv)f_{\chi^2}(v)dv}{\int f_{\chi^2}(c_{ab} + bv)f_{\chi^2}(v)dv} < \frac{\int v f_{\chi^2(\lambda)}(c_{ab} + bv)f_{\chi^2}(v)dv}{\int f_{\chi^2(\lambda)}(c_{ab} + bv)f_{\chi^2}(v)dv}.$$

With  $b > 0$ , for this it is enough that the ratio  $r(x) = f_{\chi^2(\lambda)}(x)/f_{\chi^2}(x)$  is strictly increasing for all  $x = c_{\alpha b} + bv > 0$ , this being a well-known classical result (regarding the monotone likelihood ratio property of the non-central chi square, see Karlin and Rubin 1956).  $\square$

Let  $\kappa(\gamma, a) = P((z_1 + \gamma\delta^*)'(z_1 + \gamma\delta^*) - a(z_1 + \gamma\delta^*)'z_2 > c_{\alpha a})$ , with  $a \geq 0$  and  $c_{\alpha a}$  defined such that  $P(z_1'z_1 - az_1'z_2 > c_{\alpha a}) = \alpha$ .

*Lemma 2:* Each of the following holds: (i)  $\kappa(\gamma, a) \rightarrow 1$  as  $\gamma \rightarrow \infty$ , and (ii) at each  $a > 0$ ,  $\kappa(\gamma, a)$  is decreasing in  $a$  for all  $\gamma$  sufficiently large.

*Proof:* As  $\gamma \rightarrow \infty$ ,  $(z_1 + \gamma\delta^*)'(z_1 + \gamma\delta^*) - a(z_1 + \gamma\delta^*)'z_2 \approx (2\gamma\delta^*)'z_1 - (a\gamma\delta^*)'z_2 + \gamma^2(\delta^*)'\delta^* \stackrel{d}{=} N(\lambda, (4+a^2)\lambda)$ , with  $\lambda = \delta'\delta$  and  $\delta = \gamma\delta^*$  (as earlier). For (i), as  $\gamma \rightarrow \infty$ ,  $\kappa(\gamma, a) \approx 1 - F_{\mathcal{N}}(x)$  with standard normal cumulative distribution  $F_{\mathcal{N}}$  and  $x = (c_{\alpha a} - \lambda)/\sqrt{(4+a^2)\lambda} \rightarrow -\infty$ , hence  $\kappa(\gamma, a) \rightarrow 1$ . For (ii), for a given  $a$  we can once again use  $\kappa(\gamma, a) \approx 1 - F_{\mathcal{N}}(x)$ , and as  $a$  rises incrementally  $x (< 0)$  rises, causing  $\kappa(\gamma, a)$  to fall.  $\square$

*Proof of Theorem 1:* (i) We can write  $G_{\text{nest}} \stackrel{d}{\approx} (z_1 + \delta)'(z_1 + \delta) - bz_2'z_2$  and with  $z_1, z_2, b$  as defined earlier, in which case Lemma 1(i) implies that test power increases in  $\gamma$ . Lemma 1(ii) implies that power decreases in  $b$ , and with  $b \approx (1-\rho)/\rho$ , power increases in  $\rho$ . (ii) Write  $G_{\text{split}} \stackrel{d}{\approx} (z_1 + \delta)'(z_1 + \delta) - 2\sqrt{\rho/(1-\rho)}(z_1 + \delta)'z_2 - (1/\rho + 1)z_2'z_2$ . To show that  $G_{\text{nest}}$  is unbiased for all  $\gamma$  sufficiently large, note that here  $G_{\text{nest}}$ 's distribution is approximately that of  $(z_1 + \delta)'(z_1 + \delta) - 2\sqrt{\rho/(1-\rho)}(z_1 + \delta)z_2$ , in which case we apply Lemma 2(i) with  $a = 2\sqrt{\rho/(1-\rho)}$ , and to show that power increases in  $\rho$  (for  $\gamma$  sufficiently large), we apply Lemma 2(ii). (iii) For large  $\gamma$ ,  $G_{\text{nest}}$  is approximately  $(z_1 + \delta)'(z_1 + \delta)$ , and applying Lemma 2(ii),  $G_{\text{nest}}$  has greater power than  $G_{\text{split}}$ .  $\square$

*Proof of Theorem 2:* (i)  $G_{\text{nest}} \stackrel{d}{\approx} (z_1 + \mu_1)'(z_1 + \mu_1) - b(z_2 + \mu_2)'(z_2 + \mu_2)$  with  $z_1$  and  $z_2$  independent standard normal  $p$ -vectors. Also,  $(z_2 + \mu_2)'(z_2 + \mu_2)$  is

non-central chi square with non-centrality parameter  $\lambda = \mu_2' \mu_2 > 0$ , and as noted earlier  $F_{\chi^2(\lambda)}(x) < F_{\chi^2}(x)$ , so with the independence of  $z_1$  and  $z_2$  we find  $P(G_{\text{nest}} > c)$  is less under Assumption 4 than under Assumption 3, for each  $c$ . (ii)  $G_{\text{split}} \stackrel{d}{\approx} (z_1 + \mu_1)'(z_1 + \mu_1) - a(z_1 + \mu_1)'(z_2 + \mu_2) - b(z_2 + \mu_2)'(z_2 + \mu_2)$ . Let  $\rho$  approach 0. Then  $a \downarrow 0$  and  $b \uparrow \infty$ , in which case the term  $a(z_1 + \mu_1)'(z_2 + \mu_2)$  in the  $G_{\text{split}}$  asymptotic density has vanishing influence, while the term  $b(z_2 + \mu_2)'(z_2 + \mu_2)$  has (unboundedly) increasing influence. We then follow reasoning similar to the proof of (i). (iii) For the test  $U$ , the distribution under either Assumption 3 or 4 is non-central chi square with non-centrality parameter equal to  $\mu_1' \mu_1$ , hence  $U$  has the same power under either Assumption.  $\square$

*Proof of Theorem 3:* Using (10) and (12),  $\sum_{i \in S_1} g(z_i, \hat{\psi}_1) - g(z_i, \psi^*) \approx \frac{1}{2} n_1 (\hat{\psi}_1 - \psi^*)' F (\hat{\psi}_1 - \psi^*)$  and also, with (11) and (13),  $\sum_{i \in S_1} g(z_i, \tilde{\psi}_1) - g(z_i, \psi^*) \approx \frac{1}{2} n_1 (\tilde{\nu}_1 - \nu)' F_{\nu\nu} (\tilde{\nu}_1 - \nu)$ . Hence:

$$\begin{aligned} \sum_{i \in S_1} g(z_i, \hat{\psi}_1) - g(z_i, \tilde{\psi}_1) &\approx \\ &\frac{1}{2} n_1 \left( (\hat{\psi}_1 - \psi^*)' F (\hat{\psi}_1 - \psi^*) - (\tilde{\nu}_1 - \nu)' F_{\nu\nu} (\tilde{\nu}_1 - \nu) \right). \end{aligned}$$

Further, using (12) and (13) we get the simplification  $\sum_{i \in S_1} g(z_i, \hat{\psi}_1) - g(z_i, \tilde{\psi}_1) = \frac{1}{2} n_1 \hat{\theta}_1' ((F^{-1})_{\theta\theta})^{-1} \hat{\theta}_1$ , with  $(F^{-1})_{\theta\theta}$  the upper-right  $p \times p$  sub-matrix of  $F^{-1}$ . So far the argument generalizes (transparently) the standard approach to likelihood ratio test asymptotics (as in van der Vaart 1998, Ch. 16). Similarly we obtain  $\sum_{i \in S_2} g(z_i, \hat{\psi}_1) - g(z_i, \tilde{\psi}_1) \approx n_2 \hat{\theta}_1' ((F^{-1})_{\theta\theta})^{-1} \hat{\theta}_2 - \frac{1}{2} n_2 \hat{\theta}_1' ((F^{-1})_{\theta\theta})^{-1} \hat{\theta}_1$ . We then apply (9) and the fact that  $\sqrt{n_k}(\hat{\theta}_k - \theta_k) \xrightarrow{d} N(0, (F^{-1})_{\theta\theta})$ ,  $k = 1, 2$ , to obtain  $G_{\text{nest}} \approx G_{\text{nest}}^\dagger$  and  $G_{\text{split}} \approx G_{\text{split}}^\dagger$ .  $\square$

*Proof of Theorem 4:* Let  $Z_1 = V_{\hat{\theta}_1}^{-1/2}(\hat{\theta}_1 - \theta_1)$  and  $Z_2 = V_{\hat{\theta}_2}^{-1/2}(\hat{\theta}_2 - \theta_2)$ , with  $V_{\hat{\theta}_k}$  the covariance matrix for  $\hat{\theta}_k$ ,  $k = 1, 2$ . Let  $e_1, \dots, e_n$  be the residuals from (OLS) full-sample regression of  $y$  on  $\{x, Dx, v\}$ , with  $D$  a dummy variable = 1 for observations in the first-sub-sample, = 0 otherwise. Then

$(e_1, \dots, e_n)' = C(y_1, \dots, y_n)'$  with  $C = I_n - U(U'U)^{-1}U'$ , where  $I_n$  is the  $n \times n$  identity matrix and  $U$  is the  $n \times (2p + q)$  matrix with  $i$ -th row  $(x'_i, Dx'_i, v'_i)$ . With Schur decomposition  $C = MM'$ ,  $M$  is a full rank  $n \times (n - 2p - q)$  matrix for which  $M'M = I_{n-2p-q}$ . Let  $Z_3 = \sigma^{-1}M'(e_1, \dots, e_n)'$ . With  $Z_1, Z_2, Z_3$  mutually independent standard normal vectors, to obtain the desired numerators and denominators of the posited expressions in  $Z$  for  $G$  distributions, we can (a) establish the results in the *simple case* of orthonormal regressor second moment matrix,  $L = I$ , and then (b) appeal to invariance of the  $G$  distribution with respect to  $Z$ . Step (a) yields to straightforward algebra but our derivation is rather lengthy, hence omitted. For step (b) it suffices to show that when applying the transformation  $Z \rightarrow ZJ$ , to *simple case*  $Z$  via some invertible  $(p + q) \times (p + q)$  matrix  $J$ ,  $G$ 's distribution is unaffected. This is also straightforward algebra (details omitted).  $\square$

## REFERENCES

- Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," *2nd International Symposium on Information Theory*, B. N. Petrov and F. Csaki (eds.), Akademiai Kiado, Budapest, 267-281.
- Ashley, R. (1998), "A new technique for postsample model selection and validation," *Journal of Economic Dynamics and Control* 22, 647-665.
- Chow, G. C. (1960), "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica* 28, 591-605.
- Clark, T. E. and M. W. McCracken (2001a), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* 105, 85-110.
- Clark, T. E. and M. W. McCracken (2001b), "Forecast-Based Model Selection in the Presence of Structural Breaks," unpublished manuscript, University of Missouri-Columbia.
- Clements, M. P. and D. F. Hendry (1999), *Forecasting Non-stationary Economic Time Series*, Cambridge: Cambridge University Press.
- Diebold, F. X. and R. S. Mariano (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13, 253-263.
- Efron, B. and R. J. Tibshirani (1993), "An Introduction to the Bootstrap," New York: Chapman and Hall/CRC.
- Gilbert, S. (2001), "Sampling Schemes and Hypothesis Tests in Regression Models," manuscript, Department of Economics, Southern Illinois University - Carbondale.
- Good, P. I. (2001), *A Practical Guide to Data Analysis*, Boston: Birkhäuser.

- Imhof, J. P. (1961), "Computing the Distribution of Quadratic Forms in Normal Variables," *Biometrika* 48, 419-426.
- Inoue, A. and L. Killian (2003), "In-Sample or Out-of-Sample Tests of Predictability, Which One Should We Use?" manuscript, Department of Agricultural and Resource Economics, North Carolina State University.
- Karlin, S. and H. Rubin (1956), "Distributions possessing a monotone likelihood ratio," *Journal of the American Statistical Association* 51, 637-643.
- Johnson, N. I. and S. Kotz (1970), *Continuous Univariate Distributions - 2*, New York: Houghton Mifflin Co.
- McCracken, M. W. (1999), "Asymptotics for Out of Sample Tests of Causality," manuscript, Department of Economics, Louisiana State University.
- McQuarrie, A. D. R. and C-L Tsai (1998), *Regression and Time Series Model Selection*, New York: Imperial College Press.
- Presnell, B. and D. Boos (2004), "The In-and-Out-of-Sample (IOS) Likelihood Ratio Test for Model Misspecification," forthcoming, *Journal of the American Statistical Association*.
- Romer, D. (2001), *Advanced Macroeconomics*, 2nd ed., New York: McGraw-Hill.
- Rossi, B. (2003), "Optimal tests for nested model selection with underlying parameter instability," manuscript, Department of Economics, Duke University.
- Scheffe, H. (1959), *The Analysis of Variance*, New York: Wiley.
- Schervish, M. J. (1995), *Theory of Statistics*, New York: Springer.

- Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Society* 88, 486-494.
- Shao, J. (1996), "Bootstrap Model Selection," *Journal of the American Statistical Society* 91, 655-665.
- Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection," *Statistica Sinica* 7, 221-264.
- Stone, M. (1974), "Cross-Validation Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society, Series B*, 36, 111-147.
- Stone, M. (1977), "An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion," *Journal of the Royal Statistical Society, Series B*, 44-47.
- Stuart, A., J. K. Ord and S. Arnold (1999), *Kendall's Advanced Theory of Statistics*, Volume 2A, 6th ed., London: Arnold.
- van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press.
- Wei, C. Z. (1992), "On Predictive Least Squares Principles," *The Annals of Statistics* 20, 1-42.
- West, K. D. (1996): "Asymptotic Inference About Predictive Ability," *Econometrica* 64, 1067-1084.
- Zhang, P. (1992), "On the Distributional Properties of Model Selection Criteria," *Journal of the American Statistical Association* 87, 732-737.
- Zhang, P. (1993), "Model Selection via Multifold Cross Validation," *The Annals of Statistics* 21, 299-313.

TABLE 1: Test Distribution

p	$\alpha = 10$	$\alpha = 5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 5$	$\alpha = 1$
nest, $\rho = \frac{1}{4}$			nest, $\rho = \frac{1}{2}$			nest, $\rho = \frac{3}{4}$			
1	1.4969	2.5874	5.3310	2.0689	3.1904	5.9672	2.4301	3.5622	6.3516
2	1.8326	3.2189	6.4378	3.2189	4.6051	7.8240	4.0298	5.4161	8.6350
3	1.7373	3.3851	7.0076	4.0781	5.6607	9.2144	5.3705	6.9374	10.4722
4	1.3801	3.2901	7.2977	4.7946	6.5436	10.3836	6.5895	8.3049	12.1031
5	0.8312	3.0104	7.3992	5.4220	7.3197	11.4165	7.7349	9.5797	13.6097
6	0.1349	2.5869	7.3581	5.9870	8.0208	12.3538	8.8298	10.7910	15.0304
7	-0.6761	2.0466	7.2018	6.5050	8.6656	13.2190	9.8874	11.9552	16.3876
8	-1.5774	1.4093	6.9489	6.9861	9.2658	14.0272	10.9160	13.0827	17.6949
9	-2.5514	0.6902	6.6128	7.4370	9.8297	14.7887	11.9211	14.1807	18.9620
10	-3.5856	-0.0984	6.2038	7.8628	10.3633	15.5112	12.9069	15.2543	20.1956
nest, $\rho = \frac{4}{5}$			nest, $\rho = \frac{5}{6}$			nest, $\rho = \frac{6}{7}$			
1	2.4897	3.6237	6.4148	2.5288	3.6626	6.4540	2.5551	3.6898	6.4840
2	4.1589	5.5452	8.7636	4.2405	5.6268	8.8457	4.2969	5.6832	8.9021
3	5.5710	7.1365	10.6694	5.6970	7.2618	10.7938	5.7836	7.3480	10.8794
4	6.8629	8.5755	12.3702	7.0340	8.7452	12.5374	7.1512	8.8616	12.6533
5	8.0822	9.9228	13.9473	8.2989	10.1374	14.1589	8.4470	10.2843	14.3042
6	9.2519	11.2074	15.4393	9.5145	11.4672	15.6953	9.6938	11.6448	15.8707
7	10.3848	12.4455	16.8683	10.6937	12.7508	17.1688	10.9042	12.9594	17.3744
8	11.4892	13.6475	18.2480	11.8446	13.9987	18.5933	12.0865	14.2382	18.8295
9	12.5706	14.8204	19.5879	12.9726	15.2176	19.9782	13.2461	15.4882	20.2450
10	13.6330	15.9692	20.8948	14.0819	16.4125	21.3303	14.3869	16.7144	21.6277
split, $\rho = \frac{1}{4}$			split, $\rho = \frac{1}{2}$			split, $\rho = \frac{3}{4}$			
1	1.2767	2.4039	5.2739	1.9490	3.3038	6.7018	3.1172	5.0436	9.8420
2	1.1525	2.6152	6.0108	2.5134	4.2268	8.2064	4.4628	6.8708	12.4620
3	0.4061	2.2312	6.1224	2.5814	4.6008	9.0618	5.2176	8.0060	14.2236
4	-0.8313	1.4571	5.8797	2.3614	4.6758	9.5864	5.6604	8.7892	15.5632
5	-2.4481	0.3647	5.3788	1.9376	4.5454	9.8916	5.8976	9.3440	16.6340
6	-4.3252	-0.9967	4.6668	1.3590	4.2584	10.0332	5.9848	9.7328	17.5116
7	-6.3927	-2.5753	3.7728	0.6590	3.8458	10.0448	5.9568	9.9932	18.2392
8	-8.6069	-4.3277	2.7173	-0.1382	3.3296	9.9484	5.8356	10.1496	18.8464
9	-10.9388	-6.2219	1.5173	-1.0142	2.7262	9.7600	5.6384	10.2196	19.3528
10	-13.3677	-8.2341	0.1883	-1.9558	2.0482	9.4914	5.3772	10.2168	19.7736
split, $\rho = \frac{4}{5}$			split, $\rho = \frac{5}{6}$			split, $\rho = \frac{6}{7}$			
1	3.5795	5.7480	11.1470	3.9978	6.3882	12.3360	4.3841	6.9825	13.4379
2	5.1985	7.9045	14.1880	5.8608	8.8404	15.7596	6.4694	9.7041	17.2151
3	6.1710	9.2970	16.2750	7.0230	10.4592	18.1380	7.8022	11.5283	19.8590
4	6.8065	10.3040	17.8960	7.8228	11.6616	20.0082	8.7493	12.9066	21.9555
5	7.2220	11.0635	19.2210	8.3880	12.5970	21.5568	9.4465	13.9930	23.7069
6	7.4780	11.6440	20.3345	8.7840	13.3410	22.8756	9.9652	14.8883	25.2112
7	7.6115	12.0870	21.2845	9.0498	13.9368	24.0186	10.3460	15.6205	26.5279
8	7.6475	12.4190	22.1035	9.2124	14.4150	25.0200	10.6183	16.2274	27.6927
9	7.6025	12.6585	22.8140	9.2898	14.7954	25.9044	10.8003	16.7300	28.7329
10	7.4905	12.8210	23.4315	9.2958	15.0930	26.6892	10.9074	17.1458	29.6674

**TABLE 2: Monte Carlo, Rejection Rates**

		n = 100			n = 200		
test		$\alpha = 10$	$\alpha = 5$	$\alpha = 1$	$\alpha = 10$	$\alpha = 5$	$\alpha = 1$
$H_0$	full	0.11	0.05	0.01	0.10	0.05	0.01
	nest=1/4	0.12	0.07	0.02	0.12	0.06	0.02
	nest=1/2	0.11	0.05	0.01	0.10	0.05	0.01
	nest=3/4	0.11	0.05	0.01	0.10	0.05	0.01
	split=1/4	0.12	0.07	0.02	0.12	0.07	0.02
	split=1/2	0.13	0.06	0.02	0.10	0.05	0.01
	split=3/4	0.12	0.06	0.01	0.10	0.05	0.01
$H_{\text{hom}}$	full	0.80	0.70	0.46	0.97	0.94	0.82
	nest=1/4	0.63	0.56	0.41	0.87	0.83	0.72
	nest=1/2	0.76	0.67	0.45	0.95	0.92	0.80
	nest=3/4	0.80	0.70	0.47	0.97	0.94	0.82
	split=1/4	0.56	0.51	0.36	0.78	0.75	0.65
	split=1/2	0.63	0.56	0.37	0.80	0.77	0.65
	split=3/4	0.58	0.48	0.30	0.73	0.68	0.54
$H_{\text{het}}$	full	0.72	0.61	0.38	0.97	0.94	0.82
	nest=1/4	0.41	0.35	0.23	0.44	0.41	0.32
	nest=1/2	0.36	0.31	0.19	0.40	0.36	0.26
	nest=3/4	0.32	0.25	0.14	0.36	0.31	0.20
	split=1/4	0.42	0.37	0.26	0.45	0.43	0.35
	split=1/2	0.43	0.38	0.27	0.45	0.42	0.35
	split=3/4	0.43	0.39	0.30	0.45	0.42	0.35

**TABLE 3: Inflation Dynamics**

period	$\alpha$	$\beta$	$\gamma$
1948-1969	0.003 (0.001)	-0.038 (0.016)	0.318 (0.060)
1970-2003	0.001 (.001)	0.001 (0.008)	0.667 (0.038)

U weight	U simple	W weight	W simple	G weight	G simple	$G^\dagger$
0.90 [0.34]	2.42 [0.12]	4.75 [0.03]	15.71 [0.00]	nest -6.30 [0.97]	-21.38 [1.00]	-117.81 [1.00]
				split -19.17 [0.98]	-42.78 [1.00]	-135.15 [1.00]

Note: ( ) = standard error, [ ] = asymptotic  $p$ -value.