cemmap
centre for microdata methods and practice

A Simple Parametric Model Selection Test

Susanne M. Schennach
Daniel Wilhelm

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP30/16

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

A Simple Parametric Model Selection Test

Susanne M. Schennach*
Department of Economics, Brown University
and
Daniel Wilhelm
Department of Economics, University College London†

July 27, 2016

Abstract

We propose a simple model selection test for choosing among two parametric likelihoods which can be applied in the most general setting without any assumptions on the relation between the candidate models and the true distribution. That is, both, one or neither is allowed to be correctly specified or misspecified, they may be nested, non-nested, strictly non-nested or overlapping. Unlike in previous testing approaches, no pre-testing is needed, since in each case, the same test statistic together with a standard normal critical value can be used. The new procedure controls asymptotic size uniformly over a large class of data generating processes. We demonstrate its finite sample properties in a Monte Carlo experiment and its practical relevance in an empirical application comparing Keynesian versus new classical macroeconomic models.

Keywords: uniform size control, one-step test, sample-splitting, Vuong test

*This work was made possible in part through financial support from the National Science Foundation via grants SES-0752699 and SES-1061263/1156347, and through TeraGrid computer resources provided by the University of Texas under grant SES-070003.

†The author gratefully acknowledges financial support from a Katherine Dusak Miller Fellowship, a Wesley C. Pickard PhD Fellowship, and from the ESRC Centre for Microdata Methods and Practice at IFS (RES-589-28-0001)

1 Introduction

Model selection is an important step in most empirical work and, accordingly, there exists a vast literature devoted to this issue (Cox (1961, 1962); Atkinson (1970); Mizon and Richard (1986); Gourieroux and Monfort (1994); Chesher and Smith (1997); Smith (1992, 1997); Ramalho and Smith (2002); Andrews (1997, 1999); Andrews and Lu (2001); Hong, Preston, and Shum (2003); Kitamura (2003); Zellner (1971); Leamer (1983); Sin and White (1996). Since Akaike (1973, 1974), the Kullback-Leibler (KL) information criterion has become a popular measure for descriminating among models taking the form of parametric likelihoods, especially in the context of nested generalized linear models ("analysis of deviance"; e.g. Nelder and Wedderburn (1972), McCullagh and Nelder (1989)). One strand of the literature (Nishii (1988), Vuong (1989), Sin and White (1996), Inoue and Kilian (2006), among others) uses this criterion together with earlier ideas about embedding the model selection problem into a classical hypothesis testing framework (e.g. Hotelling (1940) and Chow (1980)). In essence, this approach uses the maximum of the likelihood function as a goodness-of-fit measure. If model A is found to have a statistically significantly larger maximum likelihood than model B, then model A is to be preferred.

In an influential paper, Vuong (1989) has established that, unfortunately, the difference between the KL information criterion (KLIC) of two competing models exhibits a wide variety of limiting distributions (normal, χ^2 or even mixtures of χ^2), depending on whether the two models are overlapping or not, or whether one of the models is correctly specified or not. As a result, using the KLIC typically requires pre-testing to establish which distribution to use for the computation of critical values for the tests. There are two reasons why the resulting two-step model selection test exhibits non-uniform behavior under the null and thus may suffer from size distortions: first, the existence of different

asymptotic distributions of the test statistic implies that size distortions can occur when models are non-nested but "close" to each other. Second, the use of a pre-test induces the well-known non-uniformity of two-step testing procedures (Leeb and Pötscher (2005)) that may also lead to size distortions. Shi (2015) seeks to address this issue by proposing a modified Vuong test for non-nested models which uniformly controls size but involves solving potentially high-dimensional optimization problems to find the appropriate critical values from a nonstandard limiting distribution.

In this paper, we instead propose a simple method that delivers a model selection criterion based on the KL discrepancy and yet only involves a test statistic that is asymptotically $N(0,1)$-distributed in all cases (nested, non-nested or overlapping), under the null that the two models fit the data equally well. Therefore, no pre-testing is required, complicated limiting distributions are entirely avoided, the test uniformly controls size, and we show in simulations that it may be significantly more powerful than Vuong's test. In fact, we provide simulation results in which the Vuong test's power is close to the test's nominal size while our test has power close to one. These advantages do come at the expense of some power loss relative to Vuong's test when the models are nested. However, our simulations suggest that this effect is small and therefore insufficient to offset the advantages of the method. In addition, our simulations suggest that neither Shi's nor our test generally dominates the other in terms of power or its ability to control size.

We test the hypothesis that two models have the same KL discrepancy to the true distribution versus one of them being smaller. In case of a rejection, the model with the smaller discrepancy is retained, otherwise the criterion suggests both models fit the data equally well. Our approach remains valid even if both models are misspecified and enables the selection of the least misspecified of the two, i.e. the model with the smallest KL discrepancy from the truth. This capability fits nicely within the context of valid

likelihood inference under potential model misspecification (White (1982)). We handle the possibility of overlapping models by devising an estimator of the KLIC that smoothly interpolates between a conventional sample-splitting scheme (e.g., Yatchew (1992), Whang and Andrews (1993)) when the competing models overlap and a conventional full-sample estimator when the models do not overlap. In this fashion, the statistic of interest is never degenerate. The relative weights of the split-sample and the full-sample statistics are governed by a regularization parameter that we choose so as to trade off power and size of the test. The optimal regularization parameter requires only estimates of variance terms and therefore is very easy to compute from a given sample. In this fashion, we avoid having to consider higher-order terms of the test's asymptotic expansion (as in Vuong (1989), or, in a different hypothesis testing context, Fan and Li (1996)). Although higher-order expansions (such as Edgeworth expansions) can, in principle, be used to address the degeneracy problem, such an approach may pose significant practical problems. For instance, a higher order analysis of likelihood functions may involve quantities that are difficult to calculate for complex forms of likelihood functions (such as when it is obtained via numerical methods and/or simulations).

Besides deriving the local asymptotic power of our test we also show that it is of correct asymptotic level uniformly over a large class of data generating processes. This is a very desirable property of a test, particularly in the model selection context, as it may be difficult to judge a priori whether competing models are "close" to each other – a case in which the Vuong test exhibits potentially very large finite sample distortions due to its non-uniform behavior under the null. We also demonstrate our procedure's small sample properties in a Monte Carlo study and illustrate its practical usefulness in testing Keynesian versus new classical macroeconomic models. Finally, we discuss how our approach may be extended in various directions such as time series data or models defined by moment conditions.

Importantly, we can also apply our sample-splitting idea to tests comparing the accuracy of forecasts (such as those made popular by Diebold and Mariano (1995)) to gain asymptotic uniform size control.

Model selection is an important step in empirical research as indicated by its vast coverage in standard statistical textbooks and in statistics courses, and the large citation count of seminal papers such as Vuong (1989). In many applications such as the one we discuss in Section 9, testing which of two models possesses a smaller KL discrepancy to the truth may be of direct interest. This is, for example, the case when the two models are observable implications from competing economic theories and the model selection test then speaks to the question which of the two theories (jointly with some distributional assumptions) is a better description of the economy. Another example is that of distinguishing different theories of voter behavior as in Shi (2015). The outcome of Diebold and Mariano (1995)-type tests of which forecasting model is more accurate are also of direct interest. In all of these examples, the model selection step is not necessarily followed by another estimation or inference step.

The proofs of all results in this paper can be found in the supplementary material.

2 Setup

In this paper, we define a model to consist of a set of probability distributions over the sample space of observed variables, indexed by a finite-dimensional parameter. For example, we subsequently use models A and B defined as

$$\mathcal{P}_A := \{P_{\theta_A} \in \mathbf{P} : \ \theta_A \in \Theta_A\},$$
$$\mathcal{P}_B := \{P_{\theta_B} \in \mathbf{P} : \ \theta_B \in \Theta_B\},$$

where $\mathbf{P}$ denotes the set of all probability measures and Θ_A and Θ_B are some finite-dimensional parameter sets. Such a set of distributions could, for example, be the set of all normal distributions indexed by their means and variances. An integral part in any model selection procedure consists of choosing a criterion which measures "closeness" of two models. We consider the KLIC here because it has a variety of convenient properties one of which being that maximum likelihood estimators of θ_A in model A, say, are known to minimize the KL distance[1] between model A and the true data generating process (White (1982)). Consequently, the so-called pseudo-true parameter value θ_A^* which maximizes the population likelihood of model A delivers a distribution $P_{\theta_A^*}$ equal to the true distribution P_0 if model A is correctly specified, and can be interpreted as the best approximating model (in terms of KL distance) in the case that model A is misspecified.

More formally, define the KL distance between two distributions P and Q,[2] or if they possess densities p and q, respectively, as

$$K(P:Q) := \int \ln\left(\frac{dP}{dQ}\right) dP = E_P\left[\ln\left(\frac{p(X)}{q(X)}\right)\right].$$

The pseudo-true value θ_A^* of a model A is then defined as the one which minimizes the KL distance between model A and the true distribution P_0, viz. $\theta_A^* := \arg\min_{\theta_A \in \Theta_A} K(P_0 : P_{\theta_A})$, and similarly for model B, $\theta_B^* := \arg\min_{\theta_B \in \Theta_B} K(P_0 : P_{\theta_B})$. Under standard conditions, (quasi-) maximum likelihood estimators consistently estimate this parameter (Akaike (1973) and Sawa (1978)). If model A is correctly specified, defined as $P_0 \in \mathcal{P}_A$, then there is a true parameter $\theta_0 \in \Theta_A$ such that $P_0 = P_{\theta_A^*} = P_{\theta_0}$. We call model B nested in model A if $\mathcal{P}_B \subset \mathcal{P}_A$, non-nested if neither model is nested in the other, overlapping if $\mathcal{P}_B \cap \mathcal{P}_A \neq \emptyset$ and non-overlapping (or strictly non-nested) otherwise.

[1] Even though the KL discrepancy is not a distance metric, we will use the two terms interchangeably.

[2] Assume that P is absolutely continuous with respect to Q. Otherwise, we define the KL distance to equal $+\infty$.

The goal of this paper is to propose a model selection test for determining the model that fits the data "better". We define a model to be better if it is closer to the true distribution in the KL sense. $P_{\theta_A^*}$ and $P_{\theta_B^*}$ are the distributions in $\mathcal{P}_A$ and $\mathcal{P}_B$ which are closest to the truth, P_0, respectively. Formally, model A is defined to be better than model B if model A's KL distance to the truth is smaller than that of model B, i.e. $K(P_0 : P_{\theta_A^*}) < K(P_0 : P_{\theta_B^*})$. If the two KL distances are equal, then we say models A and B are equivalent. The procedure proposed in the next two sections selects the better model based on performing a test of

$$H_0 : \quad K(P_0 : P_{\theta_A^*}) = K(P_0 : P_{\theta_B^*}),$$

i.e. models A and B are equivalent, against model A is better, $H_A : K(P_0 : P_{\theta_A^*}) < K(P_0 : P_{\theta_B^*})$, or model B is better, $H_B : K(P_0 : P_{\theta_A^*}) > K(P_0 : P_{\theta_B^*})$.

Before proceeding to the actual model selection test, we conclude this section with the collection of a few formal definitions. To that end, let $X_i : \Omega \mapsto \mathcal{X}$, $i = 1, 2, \ldots$, be random vectors on the probability space $(\Omega, \mathcal{F}, Q_0)$ with $\mathcal{F}$ a σ-algebra and Q_0 a probability measure on Ω. Further, suppose $\mathcal{X}$ is a Polish space $\mathcal{X}$, i.e. a complete separable metric space, and $\mathcal{B}_x$ the Borel σ-algebra on $\mathcal{X}$. Denote by μ some underlying σ-finite measure on $(\mathcal{X}, \mathcal{B}_x)$, e.g. the Lebesgue measure on $\mathcal{X} = \mathbb{R}^k$. Finally, let $\mathbf{P}$ be the set of all distributions on $\mathcal{X}$ which have a measurable density with respect to μ.

3 The Test Statistic

To motivate our proposed test statistic we first briefly describe the so-called degeneracy problem that complicates the use of existing test statistics.

Let $\theta := (\theta_A', \theta_B')' \in \Theta := \Theta_A \times \Theta_B \subset \mathbb{R}^p$, let ∇_{θ_k} denote gradient vectors with respect

to θ_k, $k = A, B$, and define the moment conditions

$$E_{P_0}\ g(X;\theta) := E_{P_0}\left[\begin{pmatrix} \nabla_{\theta_A} \ln f_A(X;\theta_A) \\ \nabla_{\theta_B} \ln f_B(X;\theta_B) \end{pmatrix}\right] = 0 \tag{1}$$

which are satisfied by the pseudo-true value $\theta^* := ({\theta_A^*}', {\theta_B^*}')'$. Let $d^* := E_{P_0}[\ln f_A(X, \theta_A^*) - \ln f_B(X, \theta_B^*)]$ be the pseudo-true log-likelihood ratio of the two models. Assume that we have an i.i.d. sample $X_1, \ldots, X_n$ from P_0 and let $\hat{g}(\theta) := \sum_{i=1}^n g(X_i;\theta)/n$. We assume that $\hat{\theta} := (\hat{\theta}_A', \hat{\theta}_B')'$ is the maximum likelihood estimator of θ^*, but in principle one could use any estimator that solves the empirical analog of (1), i.e. $\hat{g}(\hat{\theta}) = o_p(1)$, typically called a "Z-estimator"[3]. GMM and GEL estimators of θ^* are examples of such estimators.

For $k = A, B$, define the variances $\sigma_k^2 := Var_{P_0}(\ln f_k(X;\theta_k^*))$, the covariance $\sigma_{AB} := Cov_{P_0}(\ln f_A(X;\theta_A^*), \ln f_B(X;\theta_B^*))$, and the variance of the likelihood ratio $\sigma^2 := \sigma_A^2 - 2\sigma_{AB} + \sigma_B^2$. Let $\hat{d}$ be the empirical log-likelihood ratio $\hat{d} := n^{-1}\sum_{i=1}^n \ln(f_A(X_i;\hat{\theta}_A)/f_B(X_i;\hat{\theta}_B))$ and define the sample variance estimators $\hat{\sigma}_k^2$ of σ_k^2, $k = A, B$, and the covariance estimator $\hat{\sigma}_{AB}$ of σ_{AB}, i.e. $\hat{\sigma}_k^2 := n^{-1}\sum_{i=1}^n (\ln f_k(X_i;\hat{\theta}_k) - \overline{\ln f_k})^2$ where $\overline{\ln f_k} := n^{-1}\sum_{i=1}^n \ln f_k(X_i;\hat{\theta}_k)$ and similarly for $\hat{\sigma}_{AB}$. The variance of the likelihood ratio, σ^2, we then estimate by $\hat{\sigma}^2 := \hat{\sigma}_A^2 - 2\hat{\sigma}_{AB} + \hat{\sigma}_B^2$.

Define t_n to be the t-statistic for testing $H_0 : d^* = 0$, i.e. $t_n := \sqrt{n}\hat{d}/\hat{\sigma}$. This statistic is equivalent to the one Vuong (1989) proposes when the two candidate models are known to be nonnested. The t-statistic possesses a standard normal limiting distribution if $\sigma^2 > 0$. The type of degeneracy ruled out by this assumption, however, poses a standard challenge encountered in parametric model selection testing. It requires that the variance of the log-likelihood ratio evaluated at the pseudo-true values is nonzero. This condition is violated when both models A and B are observationally equivalent, i.e. when both are correctly specified which implies that (i) they must be overlapping (including the nested case) and

[3]See van der Vaart (1998, Chapter 5) for an introduction.

(ii) the truth must be an element of their intersection. Then the pseudo-true densities are identical, $f_A(\cdot;\theta_A^*) \equiv f_B(\cdot;\theta_B^*)$, which in turn implies that the variance σ^2 is zero.

The common solution in the literature has been to either assume this case away or develop a pre-test for testing whether degeneracy holds or not. See Vuong (1989), Kitamura (2000) and Kitamura (2003) for a discussion of issues related to degeneracy and pre-tests that have been suggested.

We now propose a modified version of the t-statistic that preserves the standard normal limiting distribution even when the models are observationally equivalent. There are several ways one could think of regularizing the model selection problem. The approach we present here is based on re-weighting the individual log-likelihoods, which is very simple to implement and results in desirable properties of the resulting test (see Section 5). Furthermore, the efficiency loss in the "nondegenerate" observationally distinct case seems to be small in finite samples and is, in fact, asymptotically negligible under simple conditions.

For simplicity of exposition assume that the sample size n is an even number. We propose to re-weight the individual log-likelihoods

$$\hat{\bar{d}} := \frac{1}{n}\sum_{i=1}^{n}\Big(\omega_i(\hat{\varepsilon}_n)\ln f_A(X_i;\hat{\theta}_A) - \omega_{i+1}(\hat{\varepsilon}_n)\ln f_B(X_i;\hat{\theta}_B)\Big)$$

with the weights

$$\omega_k(\hat{\varepsilon}_n) := \begin{cases} 1, & k \text{ odd} \\ 1+\hat{\varepsilon}_n, & k \text{ even} \end{cases}, \qquad k = 1,\ldots,n+1 \tag{2}$$

that depend on a possibly data-dependent, real-valued regularization parameter $\hat{\varepsilon}_n$. Straightforward algebra shows that the asymptotic variance of $\sqrt{n}\hat{\bar{d}}$ can be estimated by $\hat{\bar{\sigma}}^2$, where

$$\hat{\bar{\sigma}}^2 := (1+\hat{\varepsilon}_n)\,\hat{\sigma}^2 + \frac{\hat{\varepsilon}_n^2}{2}\left(\hat{\sigma}_A^2 + \hat{\sigma}_B^2\right).$$

With the modified estimator of d^* and its variance estimator, we can construct a new t-statistic $\tilde{t}_n$ defined as

$$\tilde{t}_n := \frac{\sqrt{n}\hat{\tilde{d}}}{\hat{\tilde{\sigma}}}.$$

If $\hat{\varepsilon}_n = 0$, then $\hat{\tilde{\sigma}} = \hat{\sigma}$ and $\hat{\tilde{d}} = \hat{d}$, and the modified and unmodified t-statistics are equivalent, i.e. $\tilde{t}_n = t_n$. Now, suppose $\hat{\varepsilon}_n \neq 0$. In the observationally distinct models case, the two statistics differ only in that some observations are weighted by $1 + \hat{\varepsilon}_n$ rather than by one. To understand how the weights $\omega_k(\hat{\varepsilon}_n)$ regularize the t-statistic in the equivalent models case, rewrite the new statistic as

$$\tilde{t}_n = \frac{\sqrt{n}(\hat{d} + \hat{\varepsilon}_n \hat{d}_{split})}{\hat{\tilde{\sigma}}}$$

with

$$\hat{d}_{split} := \frac{1}{n}\sum_{i=1}^{n/2}\left(\ln f_A(X_{2i}; \hat{\theta}_A) - \ln f_B(X_{2i-1}; \hat{\theta}_B)\right).$$

This representation shows that the numerator of $\tilde{t}_n$ is equal to a weighted sum of the conventional full-sample log-likelihood ratio $\hat{d}$ and the split-sample log-likelihood ratio $\hat{d}_{split}$ which computes the log-likelihood of model A from the odd observations and that of model B from the even observations. As the data are assumed to be i.i.d., the variance of the split-sample statistic is always nonzero regardless of whether the models are observationally distinct or equivalent. The parameter $\hat{\varepsilon}_n$ determines how much of the split-sample statistic should be added to the full-sample counterpart. Equivalent models lead to identical densities, i.e. $\ln f_A(\cdot; \theta_A^*) \equiv \ln f_B(\cdot; \theta_B^*)$ and, therefore, t_n has a degenerate distribution. The new statistic $\tilde{t}_n$, however, continues to be nondegenerate because of the split-sample term. When $\hat{\varepsilon}_n \to_p 0$ at a suitable rate,[4] the net effect of the proposed regularization approach is

[4]Notice that the assumptions of Theorem 1 below do not actually require the regularization parameter to vanish with the sample size. We only need it to be bounded in probability.

to reduce to a sample splitting device in the observationally equivalent models case, while smoothly reverting to the conventional full-sample expression as the models move away from perfect overlap.

There are multiple ways one could modify the full-sample likelihood ratio statistic so as to obtain some of the desirable properties of our proposed test such as uniform asymptotic size control and the avoidance of a pre-test. For example, one could define a t-statistic based only on the split sample statistic $\hat{d}_{split}$, as sample-splitting is a known and effective way to address degeneracy issues in test statistics (e.g., Yatchew (1992), Whang and Andrews (1993)). However, when models are non-nested such a statistic may suffer from poor power as it ignores half of the sample whereas our proposed statistic does not because it asymptotically equals the full-sample likelihood ratio statistic in that case.

Another simple alternative that may at first appear attractive would be to simply pre-test whether σ^2 is significantly different from zero and accordingly use a full sample or a split sample Vuong statistic based on the result of the pre-test. While we leave the derivation of its theoretical properties for future research, we conjecture that such a two-step testing procedure is likely to suffer from similar lack of uniformity and power loss as the two-step Vuong test.

In general, two-step approaches with a discontinuous change in the second step's test statistic likely possess poor uniformity properties. A practical consequence of this problem is that practitioners could often be in the situation that very small changes to the data could yield dramatic changes in the test's p-value, which would make it hard to access the level of confidence that the chosen model is the correct one. A smooth transition between sample splitting and no sample splitting elegantly avoids this theoretical and practical problem.

The benefit of our the regularization scheme is that the strong nonsingularity condition

$\sigma^2 > 0$ can be replaced by the following very weak condition.

Assumption 1. *For $k = A, B$, $\sigma_k^2 > 0$, $Var_{P_0}((\ln f_k(X;\theta_k^*))^2) > 0$, and $Var_{P_0}(\nabla_{\theta_k} \ln f_k(X;\theta_k^*))$ is nonsingular.*

We also need standard conditions for Z-estimators to be consistent and asymptotically normal. They can be weakened substantially, but serve as a simple basis to discuss the relevant issues in our model selection framework.

Assumption 2. *$\Theta \subset \mathbb{R}^{d_\theta}$ is compact and $\ln f_k(x;\cdot)$, $k = A, B$, are twice continuously differentiable.*

For $k = A, B$, let $\nabla_{\theta_k}^2$ denote the Hessian matrix of a function of θ_k, containing derivatives with respect to elements of θ_k.

Assumption 3. *(i) $X_1, \ldots, X_n$ is an i.i.d. sequence of random variables with common distribution $P_0 \in \mathbf{P}$. (ii) There is a unique $\theta^* \in int(\Theta)$ so that $E_{P_0} g(X;\theta^*) = 0$. (iii) $E_{P_0}[\nabla_{\theta_k}^2 \ln f_k(X;\theta_k^*)]$, $k = A, B$, are invertible.*

Assumption 3(ii) can be overly restrictive because likelihoods with a unique global maximizer may possess more than one root of the corresponding first-order conditions. This means Θ has to be chosen sufficiently small so as to exclude roots not corresponding to the global maximum. The assumption is made here to simplify the exposition. In practice, however, one may simply estimate θ_A and θ_B separately by standard maximum likelihood assuming that there is a unique global maximizer.

The remainder of Assumption 3, Assumptions 1 and 2 are not very restrictive and could be termed standard regularity conditions. We also impose some moment existence conditions on the individual likelihoods and their derivatives:

Assumption 4. *(i)* $E_{P_0}[\|\nabla_{\theta_k} \ln f_k(X, \theta_k^*)\|^{2+\delta}] < \infty$ *and* $E_{P_0}[|\ln f_k(X, \theta_k^*)|^{4+\delta}] < \infty$ *for* $k = A, B$ *and some* $\delta > 0$. *(ii) There exists a function* $\bar{F}_1(x)$ *such that* $E_{P_0}\bar{F}_1(X) < \infty$ *and, for* $j, k = A, B$, *for all* $\theta = (\theta_A', \theta_B')' \in \Theta$, *for all* $x \in \mathcal{X}$, *and for* $h(x;\theta)$ *being any of the functions* $\ln f_k(x;\theta_k)$, $vec(\nabla^2_{\theta_k} \ln f_k(x;\theta_k))$ *and* $\ln f_k(x;\theta_k)\nabla_{\theta_j} \ln f_j(x;\theta_j)$, *we have* $\|h(x;\theta)\| \leq \bar{F}_1(x)$. *(iii) There exists a function* $\bar{F}_2(x)$ *such that* $E_{P_0}[|\bar{F}_2(X)|^{2+\delta}] < \infty$ *and* $\|\nabla_{\theta_k} \ln f_k(x;\theta_k)\| \leq \bar{F}_2(x)$ *for all* $x \in \mathcal{X}$ *and* $k = A, B$.

Finally, we place restrictions on the regularization parameter. First, we define the set of positive sequences that are $O(1)$ but converge to zero only at a rate slower than $n^{-1/4}$.

Definition 1. *Let* $\mathcal{E}$ *be the set of sequences* $\{\varepsilon_n\}$ *in* $\mathbb{R}$ *such that* $\varepsilon_n > 0$ *for all* $n \geq 1$, $n^{1/4}\varepsilon_n \to \infty$, *and* $\varepsilon := \lim_{n\to\infty} \varepsilon_n < \infty$.

Assumption 5. $\hat{\varepsilon}_n$ *is a sequence of real-valued, measurable functions of* $X_1, \ldots, X_n$ *such that there exists a sequence* $\{\varepsilon_n\} \in \mathcal{E}$ *with* $|\hat{\varepsilon}_n - \varepsilon_n| = O_{P_0}(n^{-1/2})$.

Notice that this assumption allows for constant ($\hat{\varepsilon}_n \equiv \varepsilon \neq 0$), deterministic and random sequences of regularization parameters $\{\hat{\varepsilon}_n\}$ as long as they do not vanish too quickly and $\{\hat{\varepsilon}_n\}$ lies in the $n^{-1/2}$-neighborhood of some deterministic sequence $\{\varepsilon_n\}$ in $\mathcal{E}$. Intuitively, we need the condition $n^{1/4}\varepsilon_n \to \infty$ to make sure the the regularization parameter does not tend to zero too quickly, otherwise it would not have any regularizing effect (at least asymptotically).

The following theorem establishes that the regularized t-statistic is asymptotically standard normal regardless of whether the two models are observationally equivalent or not.

Theorem 1. *If Assumptions 1–5 hold, then, under* H_0, $\tilde{t}_n \to_d N(0,1)$ *and, under* $H_A \cup H_B$, $|\tilde{t}_n| \to_p \infty$.

Remark 1. *Conditional densities can be accommodated just as in Vuong (1989).*

Remark 2. *The requirement $\varepsilon_n \neq 0$ (but possibly $\varepsilon_n \to 0$) is necessary only for the limiting distribution of $\tilde{t}_n$ to be nondegenerate in the observationally equivalent case. Therefore, if it is known a priori that the two models A and B are observationally distinct (e.g. strictly non-nested), $\varepsilon_n \equiv 0$ is permitted. However, Section 5 below shows that tests based on sequences that do satisfy the requirements of $\mathcal{E}$ uniformly control size. Since observationally distinct models can be "close" to observationally equivalent in finite samples, one may want to employ nonzero sequences $\{\hat{\varepsilon}_n\}$ even in such cases.*

Remark 3. *The functional form of the weights $\omega_k(\varepsilon)$ in (2) can be seen as a normalization in the following sense. In Section 6, we provide a data-driven choice of $\hat{\varepsilon}_n$ that optimizes a particular power and size trade-off given the functional form of $1 + \hat{\varepsilon}_n$ for weighting the even observations. For any other functional form of the weight, say $\bar{w}_k(\hat{\varepsilon}_n)$, the optimal $\hat{\varepsilon}_n$ would then be such that $\bar{w}_k(\hat{\varepsilon}_n) = 1 + \hat{\varepsilon}_n$ as long as the range of the function $\bar{w}_k$ is large enough. On the other hand, consider choosing some constant, say c, other than 1 for weighting the odd group together with the appropriate adjustment to the standard deviation in the denominator of $\tilde{t}_n$. This modified test statistic is numerically equivalent to our test statistic when the optimal epsilon, now $c(1+\hat{\varepsilon}_n)-1$ with $\hat{\varepsilon}_n$ the optimal choice under $c = 1$, is employed.*

Our test statistic relies on assigning individual observations to two groups. Clearly, the test statistic is invariant to sample re-orderings that permute observations within the two groups, but do not re-assign observations across the two groups. In the remainder of this section, we discuss in what sense our statistic is asymptotically invariant under re-assignment of observations across groups and the impact of such re-assignments in finite samples.

We introduced our test statistic by splitting the sample into odd and even observations, which was purely for concreteness and ease of presentation. As Theorem 1 shows, the

limiting distribution of our test statistic does not depend on the definition of the two groups. In fact, any other partition of the sample into two groups yields the same asymptotic distribution. In this sense, re-ordering has no effect on the test statistic in large samples. The supplement of this paper shows that not only does every partition of the sample into two groups lead to the same asymptotic distribution, but also the random difference between two test statistics based on different assignment rules is negligible in large samples. This result requires that one partition into two groups can be constructed from the other partition by $o(n)$ re-assignments of observations across groups.

Even though this result provides a sense in which our test statistic is asymptotically invariant to re-assignment of observations across groups, one may be concerned that, in a finite sample, the invariance may not hold. One should realize, however, that our critical values account for fluctuations due to different sample orderings, so one would have to try about 100 different re-assignments of observations across groups before finding one leading to a false rejection of the null at the 99% level (and this is assuming that re-assignment is the only source of noise, which is not the case, so, in reality, even more permutations than this would have to be tried to stumble on a permutation yielding a false rejection). The fact that our critical values account for the re-assignment noise is an automatic consequence of the fact that they account for the usual sampling noise. Indeed, a re-ordered sample is just another possible random draw from the population distribution.

To check robustness of the model selection results in finite samples, the user of our test may want to report summary statistics of covariates in the two groups. Balance of such summary statistics across the two groups ensures that estimates and test results are not driven by significant (observable) differences across the two groups. In fact, one could randomly assign observations to two groups to guarantee balance not only on observable, but also on unobservable characteristics.

Splitting samples of observations into two groups is common practice in randomized control trials, and the effect of randomization, stratification, and possible imbalance on estimators and test statistics is well-understood in that literature. The same advantages and disadvantages carry over to our context of model specification tests.

4 The Model Selection Test

The results of the previous section suggest a very simple model selection procedure based on a two-sided[5] t-test: Given a nominal level $\alpha \in (0,1)$ and some finite $\hat{\varepsilon}_n$ such as the optimal choice proposed in Section 6, we compute the test statistic $\tilde{t}_n$ and compare its absolute value to the $(1-\alpha/2)$-quantile $z_{1-\alpha/2}$ from the $N(0,1)$ distribution. If $|\tilde{t}_n| > z_{1-\alpha/2}$, then reject the null that model A and B are equally close to the truth. The rejection is in favor of model A if $\tilde{t}_n > z_{1-\alpha/2}$ and in favor of model B if $\tilde{t}_n < -z_{1-\alpha/2}$. No pre-testing is necessary and, in contrast to available methods, no complicated asymptotic distributions[6] ever need to be used.

Interestingly, conditional on a given selected model, asymptotically valid confidence regions for its parameters can be readily obtained by using the first-order conditions of its likelihood maximization problem. This scheme automatically recovers the well-known "sandwich" formula for misspecification-robust estimation of the asymptotic variance (White

[5] Alternatively, one could use a one-sided t-test with obvious modifications to the procedure.

[6] The simulation of critical values from the mixture of χ^2 distributions in Vuong (1989)'s test requires the estimation of eigenvalues of a potentially large matrix which are then to be used as the mixture weights. Such estimators may be quite imprecise in small samples and can induce further distortions. Shi (2015)'s test, on the other hand, requires some conservative critical value because the exact limiting critical value cannot be estimated consistently. The conservative critical value is then determined as the supremum over a potentially very large space of nuisance parameters which can be an expensive numerical task.

(1982), Owen (2001)). Of course, model estimation following a model selection procedure always carries the risk that the model selection step may influence the significance levels of subsequent tests. As our approach selects the best model of the two with probability approaching one, the model selection step has, asymptotically, no effect on further pointwise inference. Remark 4 below discusses uniformity properties of our procedure.

In the presence of a priori information justifying the exclusion of the observationally equivalent models case, the same test can be performed using the test statistic t_n instead of $\tilde{t}_n$. In certain modeling situations, it might be straight-forward to check whether the condition $\sigma^2 > 0$ is satisfied. For example, one might have reasons to believe that both models are only crude approximations to the truth so that both are misspecified. If, in addition, it can be established analytically that the models do not overlap, then $\sigma^2 > 0$ holds and the test without regularization can be used.

5 Large Sample Properties of the Test

5.1 Uniformity

In this section, we define a set $\mathcal{P}$ which contains all distributions under which the moment conditions and some regularity conditions similar to those in the previous section hold. Then we show that our regularized test controls size uniformly over those distributions in $\mathcal{P}$ that also satisfy the null hypothesis.

In view of the impossibility result by Bahadur and Savage (1956) and its extensions in Romano (2004), we cannot hope to gain uniform size control over general nonparametric classes of distributions. It has been recognized before (see section 11.4.2 in Lehmann and

Romano (2005), for instance) that Lyapounov's condition[7] places sufficient restrictions on the space of distributions so that one can establish uniformity for t-statistics. The following definition of the set of distributions $\mathcal{P}$ follows that route and ensures that the Lyapounov condition holds for several components of our test statistic. This can be seen as a strengthening of the assumptions in Section 3 to allow for asymptotic theory under sequences of data generating processes.

Subsequently, we need to be more specific about under which distribution P certain quantities are computed. Define $\theta^*(P) := (\theta_A^*(P)', \theta_B^*(P)')'$ to be the parameter value that satisfies $E_P g(X_i; \theta^*(P)) = 0$ and $d^*(P) := E_P[\ln f_A(X; \theta_A^*(P)) - \ln f_B(X; \theta_B^*(P))]$. Let $\sigma_k^2(P) := Var_P(\ln f_k(X; \theta_k^*(P))$, $\tilde{\sigma}^2(\theta, P, \varepsilon) := (1+\varepsilon)\sigma^2(P) + \varepsilon^2(\sigma_A^2(P) + \sigma_B^2(P))/2$, abbreviate $\tilde{\sigma}^2(\theta^*(P), P, \varepsilon)$ by $\tilde{\sigma}^2(P, \varepsilon)$, and $H_k(P) := E_P[\nabla_{\theta_k}^2 \ln f_k(X; \theta_k^*(P))])$ for $k = A, B$.

Definition 2. *For some fixed $\delta, \kappa > 0$, $0 < \underline{M} \leq \overline{M} < \infty$, and an increasing, continuous function $\epsilon : (0, \infty) \to (0, \infty)$ with $\epsilon(0) = 0$, let $\mathcal{P}$ be the set of distributions P on $\mathcal{X}$ that satisfy the following conditions for $X \sim P$: (i) There exists a unique $\theta^*(P) \in \Theta$ such that $E_P g(X; \theta^*(P)) = 0$, for all $\mu > 0$, $\inf_{\theta : \|\theta - \theta^*(P)\| \geq \mu} \|E_P g(X; \theta)\| > \epsilon(\mu)$, and $B_\kappa(\theta^*(P)) \subseteq \Theta$, where $B_\kappa(\theta)$ denotes a ball in $\mathbb{R}^{d_\theta}$ with radius κ around θ. (ii) There exists a function $D(x)$ such that $E_P[|D(X)|^{2+\delta}] \leq \overline{M}$ and, for all $x \in \mathcal{X}$,*

$$|\ln f_A(x; \theta_A^*(P)) - \ln f_B(x; \theta_B^*(P))| \leq D(x) \left(E_P \left[|\ln f_A(X; \theta_A^*(P)) - \ln f_B(X; \theta_B^*(P))|^2 \right] \right)^{1/2}, \quad (3)$$

where $\theta^*(P) := (\theta_A^*(P)', \theta_B^*(P)')'$. Further, we have $E_P[|\ln f_k(X; \theta_k^*(P))|^{4+\delta}] \leq \overline{M}$ and, similarly, $E_P[\|\nabla_{\theta_k} \ln f_k(X; \theta_k^*(P))\|^{2+\delta}] \leq \overline{M}$ for $k = A, B$. (iii) There exists a function $\bar{F}(x)$ such that $E_P \bar{F}(X) \leq \overline{M}$ and, for $j, k = A, B$, for all $\theta = (\theta_A', \theta_B')' \in \Theta$, for all $x \in \mathcal{X}$, and for $h(x; \theta)$ being any of the functions $\ln f_k(X; \theta_k)$, $\nabla_{\theta_k} \ln f_k(X; \theta_k)$, $vec(\nabla_{\theta_k}^2 \ln f_k(x; \theta_k))$

[7] See equation (23.35) in Davidson (1994), for example.

and $\ln f_k(x;\theta_k)\nabla_{\theta_j} \ln f_j(x;\theta_j)$*, we have* $\|h(x;\theta)\| \leq \bar{F}(x)$*. (iv) For* $k = A, B$*, we have* $\underline{M} \leq \lambda_{min}(H_k(P))$ *and* $\lambda_{max}(H_k(P)) \leq \overline{M}$*, where* $\lambda_{min}(A)$ *and* $\lambda_{max}(A)$*, respectively, denote the smallest and largest eigenvalue of a matrix* A*. Furthermore, for* $h(x;\theta)$ *being any of the functions* $\log f_k(x;\theta_k)$*,* $(\log f_k(x;\theta_k))^2$*, and* $\nabla_{\theta_k} \log f_k(x;\theta_k)$*,* $k = A, B$*,* $\theta := (\theta'_A, \theta'_B)'$*, we have* $\underline{M} \leq \lambda_{min}(Var(h(X;\theta^*(P))) \leq \lambda_{max}(Var(h(X;\theta^*(P))) \leq \overline{M}$*.*

Before stating the uniformity theorem, we slightly modify Assumption 5 to hold under sequences of distributions.

Assumption 6. *Let* $\hat{\varepsilon}_n$ *be a sequence of real-valued, measurable functions of* $X_1, \ldots, X_n$ *such that, for every sequence* $\{P_n\}$ *in* $\mathcal{P}$*, there exists a sequence* $\{\varepsilon_n\} \in \mathcal{E}$ *with* $|\hat{\varepsilon}_n - \varepsilon_n| = O_{P_n}(n^{-1/2})$*.*

In Section 6, we verify Assumption 6 for our proposed data-driven regularization parameter selection rule.

Theorem 2. *Suppose Assumptions 2 and 6 hold. Let* $\mathcal{P}_0 := \{P \in \mathcal{P} \colon d^*(P) = 0\}$ *be the subset of distributions in* $\mathcal{P}$ *that satisfy the null hypothesis. Then the regularized t-test of nominal level* α *is asymptotically of level* α *uniformly over* $\mathcal{P}_0$*, viz.*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} P\left(|\tilde{t}_n| > z_{1-\alpha/2}\right) = \alpha.$$

To the best of our knowledge, this uniformity property of our model selection test is the only result of this kind besides that of Shi (2015). If the test was only pointwise of correct asymptotic level, then it could be the case that for any sample size N there exists a sequence of distributions $P_n \in \mathcal{P}_0$ such that for any sample size $n \geq N$ the rejection probability under P_n is arbitrarily close to one. This possibility is ruled out when the test is uniformly of correct asymptotic level which implies that for any $\epsilon > 0$ there is a sample size N such that, for all $n \geq N$, the rejection probability under any sequence $P_n \in \mathcal{P}_0$ is at

most $\alpha + \epsilon$. Uniform control of the level over all distributions in $\mathcal{P}_0$ is both important and often difficult to establish because the distributions in the null hypothesis can be nested, non-nested or overlapping. In tests such as the Vuong test, for example, these different cases give rise to different limiting distributions of the test statistic so that even, in, say, non-nested models which are "close" to overlapping, substantial finite sample size distortions can occur. The uniformity of the level over $\mathcal{P}_0$ guarantees that such distortions do not occur or, at least, vanish in large samples. In the model selection context, this uniformity property is particularly desirable as it may be difficult to judge a priori whether competing models are "close" to each other. When they are "close", a formal model selection test is arguably the most valuable as the two models may be difficult to distinguish on other, say, theoretical grounds.

Remark 4. *Our model selection test avoids pre-testing as is necessary in Vuong's two-step procedure and guarantees uniform asymptotic size control as shown in Theorem 2. However, the well-known non-uniform behavior of post-model selection inference persists so that researchers should exercise caution when using the selected model in subsequent estimation and inference steps. In finite samples, some effect of the model selection step cannot be completely excluded (see, e.g., White (2000), Leeb and Pötscher (2005, 2008), and references therein, for a more detailed discussion). Fortunately, effective methods have been developed to quantify the effect (White (2000)).*

5.2 Local Power

Theorem 1 shows that the limiting distribution of our test statistic is independent of the regularization parameter $\hat{\varepsilon}_n$. Therefore, our test controls size (by Theorem 2 even uniformly) and is consistent against fixed alternatives, independently of the specific choice of

the sequence $\{\hat{\varepsilon}_n\}$. However, as we show in this section, the local asymptotic power of our test depends on the probability limit of $\{\hat{\varepsilon}_n\}$.

We consider local alternatives $\delta \in \mathbb{R}$ so that $n^{1/2}d^*(P_n) \to \delta$. The set $\mathcal{P}_\delta$ contains all sequences of distributions that satisfy the assumptions placed on $\mathcal{P}$ and along which $n^{1/2}d^*(P_n)$ converges to δ.

Definition 3. *For some $\delta \in \mathbb{R}$, let $\mathcal{P}_\delta$ be the set of sequences $\{P_n\}$ in $\mathcal{P}$ such that $n^{1/2}d^*(P_n) \to \delta$ and such that, for any $(\theta_{A,\infty}, \theta_{B,\infty}, \sigma_A^2, \sigma_B^2, \sigma_{AB}) \in \Theta_A \times \Theta_B \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}$, $\theta_A^*(P_n) \to \theta_{A,\infty}$, $\theta_B^*(P_n) \to \theta_{B,\infty}$, $\sigma_A^2(P_n) \to \sigma_A^2$, $\sigma_B^2(P_n) \to \sigma_B^2$, and $\sigma_{AB}(P_n) \to \sigma_{AB}$, where $\sigma_A^2(P) := Var_P(\ln f_A(X;\theta_A^*(P)))$, $\sigma_B^2(P) := Var_P(\ln f_B(X;\theta_B^*(P)))$ and $\sigma_{AB}(P) := Cov_P(\ln f_A(X;\theta_A^*(P)), \ln f_B(X;\theta_B^*(P)))$.*

Importantly, alternatives in $\mathcal{P}_\delta$ are allowed to approach both, observationally equivalent ($\sigma^2 = 0$) or observationally distinct ($\sigma^2 \neq 0$) data-generating processes, in the null. The following theorem presents the power of our test against all local alternatives in $\mathcal{P}_\delta$.

Theorem 3. *Suppose Assumptions 2 and 6 hold. Let $\{P_n\} \in \mathcal{P}_\delta$ for some localization parameter $\delta \in \mathbb{R}$. Denote by $\{\varepsilon_n\} \in \mathcal{E}$ a sequence such that $|\hat{\varepsilon}_n - \varepsilon_n| = O_{P_n}(n^{-1/2})$ and $\varepsilon := plim_{n\to\infty}\hat{\varepsilon}_n$ under P_n. Then, under P_n,*

$$\tilde{t}_n \to_d N(\tilde{\lambda}, 1)$$

with mean

$$\tilde{\lambda} := \lim_{n\to\infty} \frac{\sqrt{n}d^*(P_n)(1+\varepsilon_n/2)}{\sqrt{(1+\varepsilon_n)\sigma^2(P_n) + \varepsilon_n^2(\sigma_A^2(P_n)+\sigma_B^2(P_n))/2}},$$

and $\sigma^2(P) = \sigma_A^2(P) - 2\sigma_{AB}(P) + \sigma_B^2(P)$.

Consider sequences $\{P_n\}$ that approach an observationally distinct models case in the null, i.e. $\sigma^2(P_n) \to \sigma^2 > 0$. Then the non-centrality parameter becomes

$$\tilde{\lambda} = \frac{\delta(1+\varepsilon/2)}{\sqrt{(1+\varepsilon)\sigma^2 + \varepsilon^2(\sigma_A^2+\sigma_B^2)/2}}. \tag{4}$$

If $\{P_n\}$ approaches an equivalent models case in the null, i.e. $\sigma^2(P_n) \to 0$, and $\varepsilon \neq 0$, then

$$\tilde{\lambda} = \frac{\delta(1+\varepsilon/2)}{\varepsilon\sqrt{(\sigma_A^2+\sigma_B^2)/2}}. \tag{5}$$

In the two cases of (4) and (5), $\tilde{\lambda}$ as functions of ε is maximized at $\varepsilon = 0$ or as ε approaches 0, respectively. On the other hand, when models overlap at the truth, we require a nonzero sequence of regularization parameters, possibly converging to zero, to guarantee a nondegenerate limiting distribution of our test statistic. In finite samples, we typically encounter an intermediate case: we would prefer not to regularize ($\hat{\varepsilon}_n = 0$) if we knew that the two candidate models are "sufficiently far apart" from each other, but we would choose a positive regularization parameter when the two candidate models are "close" to overlapping to minimize size distortions.[8] The next section formalizes the trade-off between power in the distinct models case and size control in the equivalent models case, and shows how this trade-off determines an optimal regularization parameter that can easily be estimated from the data.

6 Data-driven Regularization Parameter

In this section, we provide a data-driven choice of $\hat{\varepsilon}_n$ that minimizes higher-order distortions to size and power of our test. Specifically, we balance the worse-case size distortion if the models were overlapping with the worst-case power loss if the models were not overlapping. The rationale for proceeding in this way is that, in our approach, size distortion only occurs for overlapping models while power loss only occurs when the models are not overlapping. Furthermore, in a finite sample, it may be difficult to accurately test whether the models

[8] Notice that Theorem 2 only requires a *positive* value $\hat{\varepsilon}_n$ for uniform size control, but does *not* imply that larger values $\hat{\varepsilon}_n$ lead to "better" size control in any sense.

are overlapping or not (this is the fundamental pre-testing problem we wish to avoid) and hence it is natural to consider both possibilities simultaneously. Such an approach also considerably simplifies the implementation of the method.

In the supplement to this paper, we derive an asymptotic expansion of the size of our test when the two models are overlapping, viz. for any distribution P_0 such that $d^*(P_0) = 0$ and $\sigma^2(P_0) = 0$,

$$P_0(|\tilde{t}_n| > z_{1-\alpha/2}) \leq \alpha + C_{SD}\varepsilon_n^{-1}n^{-1/2}\ln\ln n + \text{remainder}, \tag{6}$$

where C_{SD} is some constant. Similarly, we expand the power of our test when the models are non-nested, viz. we show that for sequences of local alternatives $\{P_n\}$ satisfying $d^*(P_n) = \delta n^{-1/2}$ for any given $\delta \in \mathbb{R} \setminus \{0\}$ and $\sigma^2 := \lim_{n\to\infty}\sigma^2(P_n) > 0$,

$$P_n\left(|\tilde{t}_n| > z_{1-\alpha/2}\right) = \Phi\left(z_{\alpha/2} + \frac{\delta}{\sigma}\right) + \Phi\left(z_{\alpha/2} - \frac{\delta}{\sigma}\right) - C_{PL}^*\varepsilon_n^2 + \text{remainder}, \tag{7}$$

where C_{PL}^* is some constant. Size distortion for overlapping models is decreasing in ε_n and power loss for distinct models is increasing in ε_n. Therefore, we propose a tuning parameter ε_n that balances the respective leading terms of the size distortion, i.e. the term $C_{SD}\varepsilon_n^{-1}n^{-1/2}\ln\ln n$, and power loss, i.e. the term $C_{PL}^*\varepsilon_n^2$. This tuning parameter choice can be estimated by

$$\hat{\varepsilon}_n = \left(\frac{\hat{C}_{SD}}{\hat{C}_{PL}^*}\right)^{1/3} n^{-1/6}(\ln\ln n)^{1/3} \tag{8}$$

with

$$\hat{C}_{PL}^* := \phi\left(z_{\alpha/2} - \frac{\hat{\delta}^*}{\hat{\sigma}}\right)\frac{\hat{\delta}^*(\hat{\sigma}^2 - 2(\hat{\sigma}_A^2 + \hat{\sigma}_B^2))}{4\hat{\sigma}^3}$$

$$\hat{C}_{SD} := 2\phi(z_{\alpha/2})\frac{\max\{|tr(\hat{H}_A^{-1}\hat{V}_A)|, |tr(\hat{H}_B^{-1}\hat{V}_B)|\}}{\sqrt{(\hat{\sigma}_A^2 + \hat{\sigma}_B^2)/2}}$$

estimating the constants C_{PL}^* and C_{SD}. In the expressions above, $\hat{\delta}^* := \hat{\sigma}/2(z_{\alpha/2} - \sqrt{4 + z_{\alpha/2}^2})$, $\hat{H}_k$ and $\hat{V}_k$, $k = A, B$, are estimates of $H_k := H_k(P_0)$ and $V_k := V_k(P_0)$ with $V_k(P) := E_P\left[\nabla_{\theta_k} \ln f_k\left(X_i, \theta_k^*(P)\right)\left(\nabla_{\theta_k} \ln f_k\left(X_i, \theta_k^*(P)\right)\right)'\right]$, obtained by replacing expectations by sample averages.

The proposed value of $\hat{\varepsilon}_n$ in (8) can easily be computed from the data as it requires only estimates of the matrices H_k and V_k, which have to be computed for the "sandwich" variance estimator for potentially misspecified models anyway, and the sample variances $\hat{\sigma}$, $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$.

Remark 5. *The tuning parameter $\hat{\varepsilon}_n$ in* (8) *depends on whether the models overlap or not via the dependence of C_{PL}^* on σ^2 and thus on σ_{AB}. In addition, some model-overlap-dependence is built into the test statistic itself. When the models are far from overlapping, $\hat{\varepsilon}_n$ is the prefactor of a higher-order term of the stochastic expansion of the test statistic. When models approach overlap, the leading term tends to zero and the term of next higher order (with $\hat{\varepsilon}_n$ prefactor) becomes dominant. As mentioned in footnote 8 it is worth emphasizing that Theorem 2 only requires a positive value of $\hat{\varepsilon}_n$ for any fixed n, but does not imply that larger values of $\hat{\varepsilon}_n$ lead to "better" size control in any sense.*

Remark 6. *The choice $\hat{\varepsilon}_n$ in* (8) *is derived from a particular trade-off between the worst-case size distortion if the models were overlapping with the worst-case power loss if the models were not overlapping. In principle, it would be possible to derive data-driven choices of $\hat{\varepsilon}_n$ using other criteria, such as weighted size distortion and power loss or error in rejection probability (e.g. as in Calonico, Cattaneo, and Farrell (2016)). One attractive feature of the trade-off presented here is the simplicity of the resulting choice in* (8).

7 Extensions

To simplify the presentation of our basic model selection procedure we restrict attention to a simple and stylized framework: we compare two fully specified parametric models based on the KL criterion, i.i.d. data and a t-statistic. In the supplement, we argue that our procedure applies much more generally and discuss some important, but mostly straightforward, extensions. First, one could use our test based on goodness-of-fit criteria other than KL distance. An important example would be comparing the accuracy of competing forecasts as in Diebold and Mariano (1995). Second, the limiting distribution of our test statistic requires only asymptotic normality of certain sample averages, so extensions to stationary data are straightforward. Third, instead of Z-estimators one could readily extend our test statistic to the comparison of models defined by moment conditions that can be estimated by GMM. Fourth, we could use our test to rank more than two models by incorporating it into a multiple testing framework in the usual way (e.g. Lehmann and Romano (2005) and Romano, Shaikh, and Wolf (2010)). To see this, notice that our test for the comparison of two models is simply a t-test for whether a mean, i.e. the KL discrepancy between the two models, is equal to zero or not. Ranking several models therefore requires testing whether multiple means, i.e. the KL discrepancies between all possible pairs of models, are equal to zero or not. A simple procedure that accounts for the multiplicity of hypotheses by, say, controlling the family-wise error rate, is based on individual t-tests with adjusted critical values. Examples of adjustments are Bonferoni's and Holm (1979)'s procedures, but more sophisticated step-up or step-down procedures could be used. See, for instance, Lehmann and Romano (2005) and Romano, Shaikh, and Wolf (2010) for more details.

The idea of altering a test statistic so that it preserves a normal distribution in all cases can be exploited in other contexts. In fact, since this paper was first circulated, Hsu and

Shi (2013) has considered the selection among conditional moment inequality models and argues that an effect similar to sample splitting can be accomplished by adding a generated independent normal noise to a non-normal statistic, to obtain a test statistic that is always normally distributed.

8 Simulations

This section reports Monte Carlo simulation results for two pairs of models (additional models are considered in the supplementaty material).

All simulations are based on $1,000$ Monte Carlo samples. Our test based on the regularized statistic $\tilde{t}_n$ is compared to the two-step Vuong procedure (see p. 321 in Vuong (1989)) and to Shi (2015)'s modified Vuong test.[9] We consider our test statistic for various choices of the regularization parameter: $\varepsilon_n = 0$ ("no reg"), $\varepsilon_n = 0.5$, $\varepsilon_n = 1$, and the optimal $\hat{\varepsilon}_n$ as defined in (8). The two-step Vuong procedure for a level-α test is implemented by setting the level equal to α in both individual steps.

Example 1 (Joint Normal Location)**.** *This example is similar to one of Shi (2015)'s who constructed it in order to illustrate the potentially poor power of Vuong's test. We let* $P_0 := N((0,\mu),(25,1)I)$ *where* I *is the identity matrix,* $\mathcal{P}_A := \{N((\mu_A,0),I) : \mu_A \in \Theta_A\}$*, and* $\mathcal{P}_B := \{N((0,\mu_B),I) : \mu_B \in \Theta_B\}$*. The null and alternative models are generated by varying* μ *in* $[0,2.5]$*.* $\mu = 0$ *corresponds to the null hypothesis (*$d = 0$*) and values in* $(0,2.5]$ *to alternatives* $d = \mu^2/2$*. Notice that the two models are observationally equivalent under the null, but misspecified.*

[9] Shi (2015) also compares her test to ours but does not use the optimal regularization parameter selection rule described in the present version of the paper.

Example 2 (Nonnested Regressions)**.** *This example is similar to one of Shi (2015)'s who constructed it in order to illustrate the potentially poor size control of Vuong's test. Let the random vector* $(Y_i, W_{i1}, \ldots, W_{i10})$*,* $i = 1, \ldots, n$*, satisfy the regression equation* $Y_i = 1 + \frac{\tau}{\sqrt{9}} \sum_{k=1}^{9} W_{ik} + \tau W_{i10} + \varepsilon_i$*, with* $\varepsilon_i \sim N(0, 2^2)$ *and* $(W_{i1}, \ldots, W_{i14}) \sim N(0, I)$*. Consider model A,* $Y_i = \alpha_0 + \sum_{k=1}^{9} \alpha_k W_{ik} + \varepsilon_i$ *with* $\varepsilon_i \sim N(0, \sigma_A^2)$*, and model B,* $Y_i = \beta_0 + \beta_1 W_{i10} + \varepsilon_i$ *with* $\varepsilon_i \sim N(0, \sigma_B^2)$*. For any value of* $\tau \neq 0$*, the two models have the same distance to the true model, but are both misspecified. We vary* τ *in* $[0, 2]$*.*

In Example 1, we estimate means and variances with the sample means and variances and, in Example 2, we estimate the regressions by ordinary least-squares. Notice that these estimators are just the maximum-likelihood estimators in the particular models considered here. In both examples, it is straightforward to verify the assumptions of our theoretical results in the preceding sections.

Table 1 reports the finite sample size of the different tests. In Example 2, we consider a family of null hypotheses whereas, in Example 1, we study the properties of our test as the true distance $|d^*|$ increases from zero (the null hypothesis) to a range of positive values (alternatives). Figure 1 shows the power curves for Example 1 in panels (a)–(c) and the null rejection probabilities for Example 2 in panel (d). In both examples, we report results for 5%-level tests. In addition, we also show power results at the 1% level in Example 1. The black horizontal lines in the power and size graphs mark the level of the tests. 'no reg', '$\hat{\varepsilon}_n = 0.5$', '$\hat{\varepsilon}_n = 1$', and 'optimal' refer to our test using $\hat{\varepsilon}_n = 0$, $\hat{\varepsilon}_n = 0.5$, $\hat{\varepsilon}_n = 1$, and the optimal epsilon defined in (8), respectively.

The two main findings from this simulation experiment can be summarized as follows. (i) In Table 1 and Figure 1(d), we see that all three tests control size well with our test having size very close to nominal size in most examples. Vuong's and Shi's test, on the other hand, more frequently have size well below nominal size. (ii) Our new test and Shi's

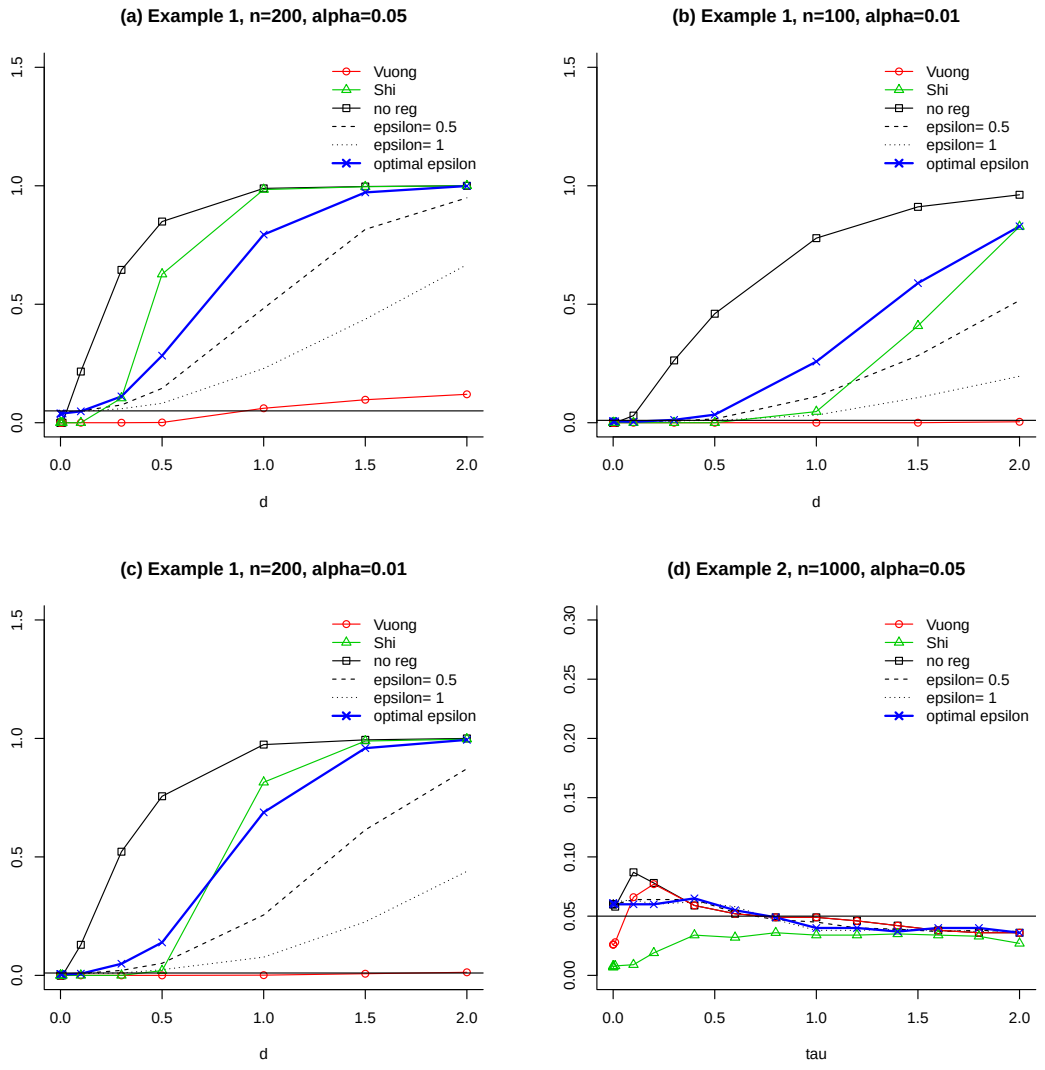

Figure 1: Comparison of the rejection frequencies of the different tests considered. For Example 1, panels (a)-(c) report power curves for different confidence levels α and sample sizes n as function of the alternative model, indexed by d. For Example 2, Panel (d) reports the actual size for a family of model pairs (indexed by τ) satisfying the null hypothesis. On all graphs, the nominal level is marked by a black horizontal line.

	our test					
n	no reg	$\varepsilon_n = 0.5$	$\varepsilon_n = 1$	optimal	Vuong	Shi
100	0.000	0.041	0.045	0.037	0.000	0.000
200	0.000	0.046	0.045	0.039	0.000	0.000
500	0.000	0.039	0.037	0.038	0.000	0.000

Table 1: Null rejection probabilities (nominal size 0.05) for Example 1.

test can have significantly higher power than Vuong's test. Since our test has size closer to nominal size than Shi's, ours possesses more power to detect alternatives close to the null, i.e. models that are difficult to distinguish. For alternatives further away from the null, neither test seems to dominate the other.

These simulations suggest that our test performs well in practice, with performance comparable and sometimes superior to existing methods. These results are especially encouraging in light of our method's conveniently straightforward implementation.

9 Empirical Application

A major part of the classic debate over (New) Keynesian versus (new) classical macroeconomic theory has focused on whether government policies, monetary or fiscal, can have any systematic impact on outcomes such as output or unemployment (Dadkhah (2009) gives a nice general overview of the literature and how it has evolved more recently). Under the new classical hypothesis of rational expectations ("RE") and natural rate of unemployment ("NR"), it has been shown (Sargent and Wallace (1975)) that, under certain assumptions, there is no such effect. Consequently, a lot of effort has been devoted to testing the joint

NR/RE hypothesis. In an influential paper, Barro (1977) proposes such a test based on a two equation system, one for money growth (DM_t),

$$DM_t = Z_t'\theta_1 + \varepsilon_{1t} \tag{9}$$

and one for unemployment (UN_t),

$$UN_t = X_t'\theta_2 + \varepsilon_{2t} \tag{10}$$

where X_t and Z_t are exogenous explanatory variables known at time $t-1$. Specifically, he suggests the covariates $Z_t := (1, DM_{t-1}, DM_{t-2}, FEDV_t, UN_{t-1})$ and $X_t := (1, DMR_t, DMR_{t-1}, DMR_{t-2}, MIL_t, MINW_t)$ with $FEDV_t$ a measure of federal government expenditure, $DMR_t := \varepsilon_{1t}$ the unanticipated part of DM_t, MIL_t a measure of military conscription and $MINW_t$ a minimum wage variable.[10] The NR/RE hypothesis implies that unemployment deviates from its so-called natural level (here proxied by MIL_t and $MINW_t$) only due to unanticipated changes in money growth (DMR_t, DMR_{t-1}, DMR_{t-2}). Therefore, equation (10) fitting the data well Barro interprets as evidence supporting the NR/RE hypothesis.

Pesaran (1982) criticizes this approach arguing that failing to reject the NR/RE hypothesis in a particular model is necessary, but not sufficient for failing to reject it against rival hypotheses. Therefore, he proposes to test it against "proper" or "genuine" alternatives, in particular against three different models with Keynesian features that satisfy (9)

[10]For exact definitions of the variables involved, see Barro (1977). He also studies output, but we confine our discussion here to unemployment as the outcome of interest.

and (10) with the following set of covariates:

$$
\begin{aligned}
K1: &\qquad X_t := (1, DM_t, DM_{t-1}, DG_t, MIL_t, MINW_t, t),\\
K2: &\qquad X_t := (1, DM_t, DM_{t-1}, DM_{t-2}, DG_t, MIL_t, MINW_t, t),\\
K3: &\qquad X_t := (1, DM_t, DM_{t-1}, DMR_t, DG_t, MIL_t, MINW_t, t),
\end{aligned}
$$

where DG_t is a measure of government spending. Subsequently, we test each of these models against Barro's new classical model and a slight variant with a time trend in the unemployment equation:

$$
\begin{aligned}
B1: &\qquad X_t := (1, DMR_t, DMR_{t-1}, DMR_{t-2}, MIL_t, MINW_t),\\
B2: &\qquad X_t := (1, DMR_t, DMR_{t-1}, DMR_{t-2}, MIL_t, MINW_t, t).
\end{aligned}
$$

We refer the reader to Pesaran (1982) for specifics about these five models and their theoretical foundations.

Based on Barro (1977)'s annual data from 1946 to 1973, we estimate each of the models in two different ways. First, we estimate both equations (9) and (10) jointly by full-information maximum likelihood (FIML) assuming that the errors in the two equations are jointly normal. Second, we estimated only the unemployment equation (10) by maximum likelihood, again assuming normality of the errors and taking the estimated series $\{DMR_t\}$ from Barro (1977) as given.

The results of the pairwise model selection tests of new classical models versus Keynesian models are reported in Table 2 and are based on the estimated optimal epsilon-parameters which ranged from 1.1 to 1.4 across the twelve pairs of models. As a sensitivity analysis we also performed our test for epsilon values in a range from 0.1 to 2.0 but the conclusions derived from the optimal epsilon do not change. When we compare Keynesian and new classical models based only the unemployment equation, all three tests fail to reject the

		K1	K2	K3
both equations	B1	-0.136	-0.664	-0.126
	B2	0.767	0.186	0.775
only unemployment equation	B1	-0.527	-1.070	-0.507
	B2	0.390	-0.247	0.408

Table 2: Value of our regularized model selection test statistic $\tilde{t}_n$ based on the optimal $\hat{\varepsilon}_n$.

hypothesis that the models are equally distant from the truth. Even adding the money growth equation does not lead to rejections. The sign of our test static suggests that the Keynesian models are closer to the truth than the new classical model B1, but further away from the the truth than B2. However, none of these statements is statistically significant at reasonable levels of confidence. Since, in the simulations, our new test tends to reject at a higher rate, both, under the null and under alternatives, with significantly higher power in some scenarios, the fact that our test fails to reject in all 12 model comparisons strenghtens the findings of the Vuong test which we found to also fail to reject in all twelve comparisons. The Vuong test's failure to distinguish the two theories is therefore less likely to be due to it under-rejecting under the null or to its potentially low power. In conclusion, we interpret the findings as there not being enough information in the present dataset to discriminate between the candidate new classical and Keynesian models. A larger sample or imposing more structure on the models might lead to different conclusions.

There are some interesting differences in these findings compared to the results reported in Pesaran (1982). He compares models based only on the unemployment equation employing an F-test as well as a Cox-type test for non-nested models. In the latter testing procedure, the null hypothesis is that model A is the true data generating process to be

tested against the alternative that model B is the truth. In terms of the F-test, no model in $\{B1, B2\}$ is found to be superior to any model in $\{K1, K2, K3\}$. His application of the Cox-type test, however, results in any model in $\{B1, B2\}$ being rejected against any alternative in $\{K1, K2, K3\}$ and vice versa. The testing outcomes of the Cox-type procedure are not possible in our test because both models are treated symmetrically: As soon as our test rejects equivalence between any two models, the one with the smaller KL distance to the truth is concluded superior to the other. Even though the null hypothesis in our test does not assume correct specification of any model, we still do not reject any model combination. Small (1979) and Pesaran (1982) criticize Barro's specification of the model and argue that the estimates of the unemployment equation may be sensitive to variations in the specification of the money growth equation. Our test results show that, at least based on the present data set, the inclusion the money growth equation has no implications on whether the new classical or the Keynesian theory is superior to the other.

SUPPLEMENTARY MATERIAL

This supplement provides the proofs of all results in the main text, additional results referenced in the main text, and additional simulations.

References

Akaike, H. (1973): "Information Theory and an Extension of the Likelihood Principle," in *Proceedings of the Second International Symposium of Information Theory*, ed. by B. N. Petrov, and F. Csáki.

Akaike, H. (1974): "A New Look at the Statistical Model Identification," in *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723.

Andrews, D. W. K. (1997): "A Conditional Kolmogorov Test," *Econometrica*, 65(5), 1097–1128.

——— (1999): "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," *Econometrica*, 67(3), 543–564.

Andrews, D. W. K., and B. Lu (2001): "Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models," *Journal of Econometrics*, 101, 123–164.

Atkinson, A. C. (1970): "A Method for Discriminating Between Models," *Journal of the Royal Statistical Society: Series B*, 32(3), 323–353.

Bahadur, R. R., and L. J. Savage (1956): "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *The Annals of Mathematical Statistics*, 27(4), 1115–1122.

Barro, R. J. (1977): "Unanticipated Money Growth and Unemployment in the United States," *The American Economic Review*, 67(2), 101–115.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2016): "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," Discussion paper.

Chesher, A., and R. J. Smith (1997): "Likelihood Ratio Specification Tests," *Econometrica*, 65(3), 627–646.

Chow, G. C. (1980): "The Selection of Variables for Use in Prediction: A Generalization of Hotelling's Solution," in *Quantitative econometrics and development*, ed. by L. Klein, M. Nerlove, and S. C. Tsiang, pp. 105–114. Academic Press, New York.

Cox, D. R. (1961): "Tests of Separate Families of Hypotheses," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 105–123. University of California Press, Berkeley.

——— (1962): "Further Results on Tests of Separate Families of Hypotheses," *Journal of the Royal Statistical Society: Series B*, 24(2), 406–424.

Dadkhah, K. (2009): *The Evolution of Macroeconomic Theory and Policy*. Springer, Berlin.

Davidson, J. (1994): *Stochastic Limit Theory*. Oxford University Press.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13(3), 253–263.

Fan, Y., and Q. Li (1996): "Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms," *Econometrica*, 64(4), 865–890.

Gourieroux, C., and A. Monfort (1994): "Testing Non-Nested Hypotheses," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. IV, pp. 2583–2637. Elsevier Science B.V.

Holm, S. (1979): "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 6, 65–70.

Hong, H., B. Preston, and M. Shum (2003): "Generalized Empirical Likelihood Based Model Selection Criteria For Moment Condition Models," *Econometric Theory*, 19(06), 923–943.

Hotelling, H. (1940): "The Selection of Variates for Use in Prediction with Some Comments on the General Problem of Nuisance Parameters," *The Annals of Mathematical Statistics*, 11(3), 271–283.

Hsu, Y.-C., and X. Shi (2013): "Model Selection Tests for Conditional Moment Inequality Models," Discussion paper, University of Wisconsin-Madison.

Inoue, A., and L. Kilian (2006): "On the Selection of Forecasting Models," *Journal of Econometrics*, 130(2), 273–306.

Kitamura, Y. (2000): "Comparing Misspecified Dynamic Econometric Models Using Nonparametric Likelihood," Discussion paper, University of Pennsylvania.

——— (2003): "A Likelihood-Based Approach to the Analysis of a Class of Nested and Non-Nested Models," Discussion paper, Yale University.

Leamer, E. E. (1983): "Model Choice and Specification Analysis," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. D. Intriligator, vol. I, pp. 285–330. North-Holland, Amsterdam.

Leeb, H., and B. M. Pötscher (2005): "Model Selection and Inference: Facts and Fiction," *Econometric Theory*, 21, 21–59.

——— (2008): "Can One Estimate the Conditional Distribution of Post Model Selection Estimators?," *Econometric Theory*, 24, 338–376.

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

McCullagh, P., and J. A. Nelder (1989): *Generalized Linear Models*. Chapman & Hall/CRC, 2nd edn.

Mizon, G. E., and J.-F. Richard (1986): "The Encompassing Principle and its Application to Testing Non-Nested Hypotheses," *Econometrica*, 54(3), 657–678.

Nelder, J. A., and R. W. M. Wedderburn (1972): "Generalized Linear Models," *Journal of the Royal Statistical Society. Series A (General)*, 135(3), 370–384.

Nishii, R. (1988): "Maximum Likelihood Principle and Model Selection when the True Model is Unspecified," *Journal of Multivariate Analysis*, 27(2), 392–403.

Owen, A. B. (2001): *Empirical Likelihood*. Chapman & Hall/CRC, New York.

Pesaran, M. H. (1982): "A Critique of the Proposed Tests of the Natural Rate-Rational Expectations Hypothesis," *The Economic Journal*, 92(367), 529–554.

Ramalho, J. J. S., and R. J. Smith (2002): "Generalized Empirical Likelihood Non-Nested Tests," *Journal of Econometrics*, 107(1-2), 99–125.

Romano, J. P. (2004): "On Non-parametric Testing, the Uniform Behaviour of the t-test, and Related Problems," *Scandinavian Journal of Statistics*, 31(4), 567–584.

Romano, J. P., A. M. Shaikh, and M. Wolf (2010): "Hypothesis Testing in Econometrics," *Annual Review of Economics*, 2(1), 75–104.

Sargent, T. J., and N. Wallace (1975): "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *The Journal of Political Economy*, 83(2), 241–254.

Sawa, T. (1978): "Information Criteria for Discriminating Among Alternative Regression Models," *Econometrica*, 46(6), 1273–1291.

Shi, X. (2015): "A Nondegenerate Vuong Test," *Quantitative Economics*, 6, 85–121.

Sin, C.-Y., and H. White (1996): "Information Criteria for Selecting Possibly Misspecified Parametric Models," *Journal of Econometrics*, 71(1-2), 207–225.

Small, D. H. (1979): "Unanticipated Money Growth and Unemployment in the United States: Comment," *The American Economic Review*, 69(5), 996–1003.

Smith, R. J. (1992): "Non-Nested Tests for Competing Models Estimated by Generalized Method of Moments," *Econometrica*, 60(4), 973–980.

——— (1997): "Alternative Semi-Parametric Likelihood Approaches to Generalised Method of Moments Estimation," *The Economic Journal*, 107(441), 503–519.

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), 307–333.

Whang, Y.-J., and D. W. K. Andrews (1993): "Tests of Specification for Parametric and Semiparametric Models," *Journal of Econometrics*, 57, 277–318.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50(1), 1–25.

——— (2000): "A Reality Check for Data Snooping," *Econometrica*, 68(5), 1097–1126.

YATCHEW, A. J. (1992): "Nonparametric Regression Tests Based on Least Squares," *Econometric Theory*, 8, 435–451.

ZELLNER, A. (1971): *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.