

Schennach, Susanne; Wilhelm, Daniel

**Working Paper**

## A simple parametric model selection test

cemmap working paper, No. CWP10/14

**Provided in Cooperation with:**

The Institute for Fiscal Studies (IFS), London

*Suggested Citation:* Schennach, Susanne; Wilhelm, Daniel (2014) : A simple parametric model selection test, cemmap working paper, No. CWP10/14, Centre for Microdata Methods and Practice (cemmap), London,  
<https://doi.org/10.1920/wp.cem.2014.1014>

This Version is available at:

<https://hdl.handle.net/10419/97390>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

# A simple parametric model selection test

---

**Susanne Schennach**  
**Daniel Wilhelm**

The Institute for Fiscal Studies  
Department of Economics, UCL

**cemmap** working paper CWP10/14

# A Simple Parametric Model Selection Test\*

Susanne M. Schennach<sup>†</sup>  
Brown University

Daniel Wilhelm<sup>‡</sup>  
UCL and CeMMAP

March 12, 2014

## 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.

---

\*First version: August 14, 2009

<sup>†</sup>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. E-Mail: [smschenn@brown.edu](mailto:smschenn@brown.edu)

<sup>‡</sup>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). E-Mail: [d.wilhelm@ucl.ac.uk](mailto:d.wilhelm@ucl.ac.uk)

# 1 Introduction

Model selection is an important step in most empirical work and, accordingly, there exists a vast literature devoted to this issue. Since Akaike (1973, 1974), the Kullback-Leibler (KL) information criterion has become a popular measure for discriminating among models taking the form of parametric likelihoods. 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,  $\chi^2$  or even mixtures of  $\chi^2$ ), 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. This situation considerably complicates the model selection procedure. Even if it is easy to analytically establish that competing models are non-nested, the two may be close in terms of KLIC so that the finite sample distribution of model selection tests can suffer from distortions. Therefore, it is desirable to have a model selection procedure that does not require the knowledge of whether the models are nested, non-nested or overlapping.

To address this issue, we introduce 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 of the cases mentioned above, under the null that the two models fit the data equally well. Therefore, no pre-testing is required and complicated limiting distributions are entirely avoided. This advantage does come at the expense of some occasional power loss, but our results indicate that the extent of this effect seems insufficient to offset the advantages of the method. The general idea is to cast the model selection problem as a set of overidentifying moment restrictions that can be tested within a general M- or Z-estimation framework (e.g. GMM). Specifically, 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. 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 \(1996\)](#)).

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 process. 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. 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.

Numerous alternative approaches to model selection are well known in the literature, for example methods based on Cox tests ([Cox \(1961, 1962\)](#)) for non-nested hypotheses which generalize the likelihood ratio tests for nested hypotheses. Also, [Atkinson \(1970\)](#) proposes to nest competing models into a larger model and then use standard tests for nested hypotheses. [Mizon and Richard \(1986\)](#) suggest the use of a closely related concept, the so-called encompassing principle, as a model building device that unifies the nested and non-nested approaches. [Gourieroux and Monfort \(1994\)](#) provide an excellent survey of this large stream of the literature. Notable more recent contributions include the specification tests proposed by [Chesher and Smith \(1997\)](#), the non-nested tests in moment condition frameworks by [Smith \(1992, 1997\)](#) and [Ramalho and Smith \(2002\)](#), the conditional Kolmogorov test of [Andrews \(1997\)](#) and the moment and model selection procedures by [Andrews \(1999\)](#), [Andrews and Lu \(2001\)](#), [Hong, Preston, and Shum \(2003\)](#) and [Kitamura \(2003\)](#).

Apart from hypothesis testing, another popular approach to model selection is to embed the problem into a decision-theoretic framework and to specify a loss function over models. The model implying the smallest loss is found either from a Bayesian perspective by updating prior knowledge about model space (see [Zellner \(1971\)](#) for a comprehensive treatment of this idea) or based on model selection criteria like the Akaike Information Criterion, for instance, which trade off how well a model fits the data and its complexity (see [Leamer \(1983\)](#) for a survey of this approach). [Sin and White \(1996\)](#) more recently show how information criteria can be employed to select among possibly nested, non-nested or overlapping as well as potentially misspecified models.

Finally, [Shi \(2013\)](#) complements our approach in an interesting way. She proposes a modified Vuong test for non-nested models which uniformly controls size. While our test possesses similar uniformity properties, our main goals are to propose a test statistic for model selection whose limiting distribution is the same irrespectively of whether the competing models are nested, non-nested, overlapping, correctly specified or misspecified. In consequence, unlike [Shi \(2013\)](#) we do not require solving potentially high-dimensional optimization problems to find critical values. Our simulations suggest that neither Shi’s nor our test generally dominates the other in terms of power or its ability to control size.

Since the first draft of this paper, the idea of altering a model selection test statistic so that it preserves

a normal distribution in all cases has been exploited in related contexts. More specifically, [Hsu and Shi \(2013\)](#) considers 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.

The next section describes the model selection framework in which [Section 3](#) and [4](#) introduce our new test statistic and model selection test, respectively. [Section 5](#) derives the local asymptotic power and uniformity properties of the test for general, possibly random, regularization parameters. [Section 6](#) then derives the optimal regularization parameter that trades off power and size of the resulting test. [Section 7](#) discusses straightforward extensions to time series models and moment condition models, among other useful generalizations. [Sections 8](#) and [9](#) report a Monte Carlo study and the empirical application in which our model selection procedure’s practical relevance is demonstrated. [Section 10](#) concludes and the appendix contains all mathematical proofs.

## 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

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

where  $\mathbf{P}$  denotes the set of all probability measures and  $\Theta_A$  and  $\Theta_B$  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  $\theta_A$  in model A, say, are known to minimize the KL distance<sup>1</sup> between model A and the true data generating process ([White \(1982\)](#)). Consequently, the so-called pseudo-true parameter value  $\theta_A^*$  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  $Q$ ,<sup>2</sup> 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  $\theta_A^*$  of a model A is then defined as the one which minimizes the KL distance

---

<sup>1</sup>Even though the KL discrepancy is not a distance metric, we will use the two terms interchangeably.

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

between model A and the true distribution  $P_0$ , viz.

$$\theta_A^* := \arg \min_{\theta_A \in \Theta_A} K(P_0 : P_{\theta_A}), \quad (1)$$

and similarly for model B,

$$\theta_B^* := \arg \min_{\theta_B \in \Theta_B} K(P_0 : P_{\theta_B}). \quad (2)$$

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  $P_0 = P_{\theta_0^*} = P_{\theta_0}$ . 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.

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

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

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  $X_i : \Omega \mapsto \mathcal{X}$ ,  $i = 1, 2, \dots$ , be random vectors on the probability space  $(\Omega, \mathcal{F}, Q_0)$  with  $\mathcal{F}$  a  $\sigma$ -algebra and  $Q_0$  a probability measure on  $\Omega$ . Further, suppose  $\mathcal{X}$  is a Polish space  $\mathcal{X}$ , i.e. a complete separable metric space, and  $\mathcal{B}_x$  the Borel  $\sigma$ -algebra on  $\mathcal{X}$ . Denote by  $\mu$  some underlying  $\sigma$ -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  $\mu$ .

### 3 The Test Statistic

To simplify the presentation of our test, we first consider the case of what we call observationally distinct models whose assumptions rule out the overlapping models case. In the second part of this section, we then show how the test can be generalized to transparently handle both the observationally distinct and the overlapping case.

Let  $\theta := (\theta'_A, \theta'_B)' \in \Theta := \Theta_A \times \Theta_B \subset \mathbb{R}^p$ , let  $\nabla_{\theta_k}$  denote gradient vectors with respect to  $\theta_k$ ,  $k = A, B$ , and define the moment conditions

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

which are satisfied by the pseudo-true value  $\theta^* := (\theta_A^*, \theta_B^*)'$  as defined in (1) and (2). 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  $X_1, \dots, X_n$  from  $P_0$  and let  $\hat{g}(\theta) := \sum_{i=1}^n g(X_i; \theta)/n$ . Let  $\hat{\theta} := (\hat{\theta}'_A, \hat{\theta}'_B)'$  be an estimator that solves the empirical analog of (3), i.e.  $\hat{g}(\hat{\theta}) = 0$ , typically called a ‘‘Z-estimator’’<sup>3</sup>. GMM, GEL, and maximum likelihood estimators of  $\theta$  are possible examples.

### 3.1 The Case of Observationally Distinct Models

The following assumptions are the standard conditions for Z-estimators to be asymptotically normal. They can be weakened substantially, but serve as a simple basis to discuss the relevant issues in our model selection framework.

**Assumption 1.**  $\Theta \subset \mathbb{R}^{d_\theta}$  is compact and  $\ln f_k(x; \cdot)$ ,  $k = A, B$ , are twice continuously differentiable in a neighborhood of  $\theta_k^*$  for all  $x \in \mathcal{X}$ .

**Assumption 2.** (i)  $X_1, \dots, X_n$  is an i.i.d. sequence of random variables with common distribution  $P_0 \in \mathbf{P}$ .

(ii) There is a unique  $\theta^* \in \text{int}(\Theta)$  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.

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

**Assumption 3.** (i)  $E_{P_0}[\sup_{\theta_k \in N_k} |\ln f_k(X; \theta_k)|^2] < \infty$ , where  $N_k$  is some neighborhood of  $\theta_k^*$ , and  $E_{P_0}[\|\nabla_{\theta_k} \ln f_k(X; \theta_k^*)\|^2] < \infty$  for  $k = A, B$ .

(ii) There exists a function  $\bar{F}(x)$  such that  $E_{P_0} \bar{F}(X) < \infty$  and, for  $k = A, B$ , for all  $\theta_k \in \Theta_k$  and for all  $x \in \mathcal{X}$ , we have (a)  $|\ln f_k(x; \theta_k)| \leq \bar{F}(x)$ , (b)  $\|\nabla_{\theta_k} \ln f_k(x; \theta_k)\| \leq \bar{F}(x)$ , and (c)  $\|\text{vec}(\nabla_{\theta_k}^2 \ln f_k(x; \theta_k))\| \leq \bar{F}(x)$ .

**Assumption 4.**  $\sigma^2 > 0$ .

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 standard variance estimators  $\hat{\sigma}_k^2$  of  $\sigma_k^2 := \text{Var}_{P_0}(\ln f_k(X; \theta_k^*))$ ,  $k = A, B$ , and the covariance estimator  $\hat{\sigma}_{AB}$  of  $\sigma_{AB} := \text{Cov}_{P_0}(\ln f_A(X; \theta_A^*), \ln f_B(X; \theta_B^*))$ , 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,  $\sigma^2 = \sigma_A^2 - 2\sigma_{AB} + \sigma_B^2$ , we then estimate by  $\hat{\sigma} := \hat{\sigma}_A^2 - 2\hat{\sigma}_{AB} + \hat{\sigma}_B^2$ .

<sup>3</sup>See van der Vaart (1998, Chapter 5) for an introduction.



Define  $t_n$  to be the t-statistic for testing  $H_0 : d^* = 0$ , i.e.

$$t_n := \frac{\sqrt{n}\hat{d}}{\hat{\sigma}}. \tag{3}$$

This statistic is equivalent to the one [Vuong \(1989\)](#) proposes when the two candidate models are known to be nonnested. All asymptotic results are for  $n \rightarrow \infty$ .

**Theorem 1.** *If Assumptions 1–4 hold, then, under  $H_0$ ,  $t_n \rightarrow_d N(0, 1)$ , and, under  $H_A \cup H_B$ ,  $|t_n| \rightarrow_p \infty$ .*

Assumption 2(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  $\Theta$  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  $\theta_A$  and  $\theta_B$  separately by standard maximum likelihood assuming that there is a unique global maximizer.

The remainder of Assumption 2 and Assumptions 1, 3 are not very restrictive and could be termed standard regularity conditions.

The type of degeneracy ruled out by Assumption 4, however, poses a standard challenge encountered in parametric model selection testing. Assumption 4 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 (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  $\sigma^2$  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.

In the next subsection, we show how to regularize the t-statistic so that the observationally equivalent case can be handled as well.

### 3.2 The General Case

We subsequently show that our proposed test statistic  $t_n$  can be slightly modified in such a way that the asymptotic results from Theorem 1 transparently extend to the observationally equivalent models case. The new statistic, called  $\tilde{t}_n$ , can be used in all possible situations, whether Assumption 4 holds or not.

There are several ways one could think of regularizing the model selection problem. The approach we present here is based on reweighting 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 reweight the individual log-likelihoods

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

with the weights

$$\omega_k(\hat{\varepsilon}_n) := \begin{cases} 1, & k \text{ odd} \\ 1 + \hat{\varepsilon}_n, & k \text{ even} \end{cases}, \quad k = 1, \dots, n+1 \quad (4)$$

that depend on a possibly data-dependent, real-valued regularization parameter  $\hat{\varepsilon}_n$ . With this modified estimator of  $d^*$  and an appropriately adjusted variance estimator

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

we can construct a new t-statistic  $\tilde{t}_n$  defined as

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

If  $\hat{\varepsilon}_n = 0$ , then  $\hat{\sigma} = \hat{\sigma}$  and  $\hat{d} = \hat{d}$ , 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

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

with

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

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,  $\hat{d} = \hat{\sigma} = 0$  and  $t_n$  has a degenerate distribution. The new statistic  $\tilde{t}_n$ , however, continues to be nondegenerate because of the split-sample term. When  $\hat{\varepsilon}_n \rightarrow_p 0$  at a suitable rate,<sup>4</sup> the net effect of the proposed regularization approach is 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.

The benefit of the regularization scheme is that the strong nonsingularity condition Assumption 4 can be replaced by the following very weak condition.

**Assumption 5.** 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.

<sup>4</sup>Notice that the assumptions of Theorem 2 below do not actually require the regularization parameter to vanish with the sample size. We only need it to be bounded in probability.

We also need to slightly strengthen Assumption 3.

**Assumption 6.** (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  $\ln f_k(x; \theta_k)$ ,  $\text{vec}(\nabla_{\theta_k}^2 \ln f_k(x; \theta_k))$  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  $n \geq 1$ ,  $n^{1/4}\varepsilon_n \rightarrow \infty$ , and  $\varepsilon := \lim_{n \rightarrow \infty} \varepsilon_n < \infty$ .

**Assumption 7.**  $\hat{\varepsilon}_n$  is a sequence of real-valued, measurable functions of  $X_1, \dots, X_n$  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  $\mathcal{E}$ .

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 2.** If Assumptions 1, 2, and 5–7 hold, then, under  $H_0$ ,  $\tilde{t}_n \rightarrow_d N(0, 1)$  and, under  $H_A \cup H_B$ ,  $|\tilde{t}_n| \rightarrow_p \infty$ .

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

**Remark 2.** The requirement  $\varepsilon_n \neq 0$  (but possibly  $\varepsilon_n \rightarrow 0$ ) 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.** For ease of presentation, we chose to split the sample into two groups by selecting odd and even observations. As Theorem 2 shows, the limiting distribution of our test statistic does not depend on this definition of the groups. In fact, any other partition of the sample into two groups would yield the same asymptotic distribution. One can even show a somewhat stronger statement, viz. that our test statistic is asymptotically equivalent to a statistic that is computed in the same way except that  $o(n)$  observations from the even group are exchanged with observations from the odd group.

**Remark 4.** *The functional form of the weights  $\omega_k(\varepsilon)$  in (4) can be seen as a normalization in the following sense. In Section 6, we provide an optimal data-driven choice of  $\hat{\varepsilon}_n$  given the functional form of  $1 + \hat{\varepsilon}_n$  for weighting the even observations. For any other functional form of the weight, say  $w(\hat{\varepsilon}_n)$ , the optimal  $\hat{\varepsilon}_n$  would then be such that  $w(\hat{\varepsilon}_n) = 1 + \hat{\varepsilon}_n$  as long as the range of the function  $w$  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.*

## 4 The Model Selection Test

The results of the previous section suggest a very simple model selection procedure based on a two-sided<sup>5</sup> t-test.

**Step 1:** Choose some nominal level  $\alpha \in (0, 1)$  and some finite  $\hat{\varepsilon}_n$  such as the optimal choice proposed in Section 6.

**Step 2:** Compute the test statistic  $\tilde{t}_n$  and compare its absolute value to the  $(1 - \alpha/2)$ -quantile  $z_{1-\alpha/2}$  of the  $N(0, 1)$  distribution.

**Step 3:** 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  $\tilde{t}_n < -z_{1-\alpha/2}$ .

Absolutely no pre-testing is necessary and, in contrast to available methods, no complicated asymptotic distributions<sup>6</sup> ever need to be used.

Interestingly, once the best model has been selected (say, A), 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 (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.<sup>7</sup> In finite samples, however, some effect cannot be completely excluded. Fortunately, effective methods have been developed to quantify the effect (White (2000)).

<sup>5</sup>Alternatively, one could use a one-sided t-test with obvious modifications to the procedure.

<sup>6</sup>The simulation of critical values from the mixture of  $\chi^2$  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 (2013)’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.

<sup>7</sup>Of course, if one allows for drifting sequences of models, then the probability of correct model selection may not approach 1 asymptotically.

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 Assumption 4 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 Assumption 4 holds and the test without regularization can be used.

## 5 Large Sample Properties of the Test

### 5.1 Uniformity

In this section, we have to be more specific about which distribution certain quantities are constructed from. 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  $\sigma_k^2(P) := \text{Var}_P(\ln f_k(X; \theta_k^*(P)))$ ,  $\tilde{\sigma}^2(\theta, P, \varepsilon) := (1 + \varepsilon)\sigma^2(P) + \varepsilon^2(\sigma_A^2(P) + \sigma_B^2(P))/2$ , 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  $k = A, B$ .

We define the set  $\mathcal{P}$  to contain all distributions under which the moment conditions and the regularity conditions from 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<sup>8</sup> 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.

**Definition 2.** For some fixed  $\delta, \kappa > 0$ , and  $0 < \underline{M} \leq \overline{M} < \infty$ , 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  $E_P g(X; \theta^*(P)) = 0$  and  $B_\kappa(\theta^*(P)) \subseteq \Theta$ , where  $B_\kappa(\theta)$  denotes a ball in  $\mathbb{R}^{d_\theta}$  with radius  $\kappa$  around  $\theta$ .

(ii) There exists a function  $D(x)$  such that  $E_P[|D(X)|^{2+\delta}] \leq \overline{M}$  and, for all  $x \in \mathcal{X}$ ,

$$|\ln f_A(x; \theta_A^*(P)) - \ln f_B(x; \theta_B^*(P))| \leq D(x) \left( E_P \left[ |\ln f_A(X; \theta_A^*(P)) - \ln f_B(X; \theta_B^*(P))|^2 \right] \right)^{1/2}, \quad (5)$$

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  $\ln f_k(X; \theta_k)$ ,  $\nabla_{\theta_k} \ln f_k(X; \theta_k)$ ,  $\text{vec}(\nabla_{\theta_k}^2 \ln f_k(x; \theta_k))$  and  $\ln f_k(x; \theta_k) \nabla_{\theta_j} \ln f_j(x; \theta_j)$ , we have  $\|h(x; \theta)\| \leq \bar{F}(x)$ .

<sup>8</sup>see, for instance, equation (23.35) in Davidson (1994)

(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  $\underline{M} \leq \text{Var}(h(X; \theta^*(P))) \leq \overline{M}$  for  $k = A, B$ .

Before stating the uniformity theorem, we slightly modify Assumption 7 to hold under sequences of distributions.

**Assumption 8.** Let  $\hat{\varepsilon}_n$  be a sequence of real-valued, measurable functions of  $X_1, \dots, X_n$  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 8 for our proposed data-driven regularization parameter selection rule.

**Theorem 3.** Suppose Assumptions 1 and 8 hold. Let  $\mathcal{P}_0$  be the subset of distributions in  $\mathcal{P}$  that satisfy the null hypothesis  $d^*(P) = 0$ . Then the regularized  $t$ -test of nominal level  $\alpha$  is uniformly asymptotically of level  $\alpha$ , viz.

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

and

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_0} P(\tilde{t}_n > z_{1-\alpha}) = \alpha.$$

This uniformity property of our test is very desirable in the model selection context and, to the best of our knowledge, the only result of this kind besides that of Shi (2013). 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.

**Remark 5.** Theorem 3 continues to hold when one replaces the dominance condition (5) by the assumption that the tuning parameter  $\varepsilon_n$  is bounded away from zero. In that case, our test statistic uniformly controls size even under sequences of distributions  $P_n$  approaching the overlapping case in such a way that second moments,  $\sigma^2(P_n)$ , vanish at a faster rate than first moments,  $d^*(P_n)$ .

## 5.2 Local Power

Theorem 2 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 3 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  $\{\hat{\varepsilon}_n\}$ .

Since sequences of alternatives approaching the null at rate  $n^{-1/2}$  are the only ones leading to non-trivial asymptotic power of our test, we consider local alternatives  $\delta$  so that  $\sqrt{nd^*(P_n)} \rightarrow \delta$ . The set

$\mathcal{P}(\delta)$  contains all sequences of distributions that satisfy the assumptions placed on  $\mathcal{P}$  and along which  $\sqrt{nd^*}(P_n)$  converges to  $\delta$ .

**Definition 3.** Let  $\mathcal{P}(\delta)$  be the set of sequences  $\{P_n\}$  in  $\mathcal{P}$  such that  $\sqrt{nd^*}(P_n) \rightarrow \delta$  and such that, for any  $(\theta_{A,\infty}, \theta_{B,\infty}, \sigma_A^2, \sigma_B^2, \sigma_{AB}) \in \Theta_A \times \Theta_B \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}$ ,  $\theta_A^*(P_n) \rightarrow \theta_{A,\infty}$ ,  $\theta_B^*(P_n) \rightarrow \theta_{B,\infty}$ ,  $\sigma_A^2(P_n) \rightarrow \sigma_A^2$ ,  $\sigma_B^2(P_n) \rightarrow \sigma_B^2$ , and  $\sigma_{AB}(P_n) \rightarrow \sigma_{AB}$ , where  $\sigma_A^2(P) := \text{Var}_P(\ln f_A(X; \theta_A^*(P)))$ ,  $\sigma_B^2(P) := \text{Var}_P(\ln f_B(X; \theta_B^*(P)))$  and  $\sigma_{AB}(P) := \text{Cov}_P(\ln f_A(X; \theta_A^*(P)), \ln f_B(X; \theta_B^*(P)))$ .

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 4.** Suppose Assumptions 1 and 8 hold. Let  $\{P_n\} \in \mathcal{P}(\delta)$  for some localization parameter  $\delta \in \mathbb{R} \cup \{-\infty, +\infty\}$  and  $\varepsilon := \text{plim}_{n \rightarrow \infty} \hat{\varepsilon}_n$  under  $P_n$ . Then, under  $P_n$ ,

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

with mean

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

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

As Figure 1 illustrates and a simple calculation based on the expression for  $\tilde{\lambda}$  confirms, the local asymptotic power is maximized at  $\varepsilon = 0$ . On the other hand, when models overlap at the truth, then 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. 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 a finite sample.

## 6 Data-driven Choice of the Regularization Parameter

Theorem 4 shows that as long as  $\hat{\varepsilon}_n$  is nonzero and converges to zero in probability under  $P_n$ , the particular choice of regularization parameter sequence has no first-order effect on the asymptotic distribution of our test statistic. However, in finite samples, we cannot rule out that the particular choice of  $\hat{\varepsilon}_n$  has an effect on the finite sample distribution of  $\tilde{t}_n$ .

In this section, we provide bounds on the higher-order size distortion and power loss of our test and derive the regularization parameter  $\hat{\varepsilon}_n$  that trades off these two worst case errors. Interestingly, it turns out to possess a very simple analytic expression. Let  $\hat{H}_k$  and  $\hat{V}_k$ ,  $k = A, B$ , be estimates of  $H_k := H_k(P_0)$  and  $V_k := V_k(P_0)$ , where  $V_k(P) := E_P[\nabla_{\theta_k} \ln f_k(X_i, \theta_k^*(P)) (\nabla_{\theta_k} \ln f_k(X_i, \theta_k^*(P)))]$ ,

obtained by replacing expectations by sample averages. Then the the optimal regularization parameter is

$$\hat{\varepsilon}_n = \hat{c}_\alpha n^{-1/4} \sqrt{\ln \ln n} \quad (6)$$

with

$$\hat{c}_\alpha := \sqrt{\frac{4\phi(z_{1-\alpha})\hat{\Lambda}}{\phi(z_b)(z_b + z_{1-\alpha})(\hat{\sigma}_A^2 + \hat{\sigma}_B^2)}} \\ \hat{\Lambda} := \max \left\{ \left| \text{tr}(\hat{H}_A^{-1}\hat{V}_A) \right|, \left| \text{tr}(\hat{H}_B^{-1}\hat{V}_B) \right| \right\}$$

and  $z_b := \frac{1}{2}(-z_{1-\alpha} + (4 + z_{1-\alpha}^2)^{1/2})$ , which can be shown to be the location of the largest power loss of our test.  $z_{1-\alpha}$  denotes the  $(1-\alpha)$ -quantile of the standard normal distribution. The proposed parameter (6) can easily be estimated 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  $\hat{\sigma}_A^2$  and  $\hat{\sigma}_B^2$ . The remainder of this section formally discusses in what sense (6) is optimal.

To allow for local alternatives in the power calculations, we consider a sequence  $P_n$  of generating processes.

**Assumption 9.** For any  $n \in \mathbb{N}$ , the  $X_{ni}$  for  $i = 1, \dots, n$  are iid random variables taking value in  $\mathcal{X}$  and drawn from the probability measure  $P_n$  converging weakly to some measure  $P_0$  and each  $P_n(x)$  admits a Radon-Nikodym derivative  $p_n(x)$  with respect to  $P_0(x)$ .

**Definition 4.** We say that  $g : \mathcal{X} \times \Theta \mapsto \mathbb{R}^{d_g}$  for  $d_g \in \mathbb{N}$  and  $\Theta$  is compact (under some metric  $d_\theta(\cdot, \cdot)$ ) satisfies a **triangular array dominance condition** if

1.  $g(x, \theta)$  is continuous in  $\theta$  at each  $(x, \theta) \in \mathcal{X} \times \Theta$ ;
2. There exists  $G(x)$  such that  $E_{P_0}[G(X_{0i})] < \infty$  (for  $X_{0i}$  drawn from  $P_0$ ) and such that, for all  $\theta \in \Theta$  and  $n \in \mathbb{N}$ ,  $\|g(x, \theta)\| p_n(x) \leq G(x)$  for all  $x \in \mathcal{X}$  and for  $p_n(x)$  as in Assumption 9;
3. There exists  $\bar{G} < \infty$  such that  $E_{P_n}[\|g(X_{ni}, \theta)\|^4] \leq \bar{G}$  for all  $i = 1, \dots, n$ , all  $n \in \mathbb{N}$  and all  $\theta \in \Theta$ .

**Assumption 10.**  $\ln f_A(x, \theta_A)$  and  $\ln f_B(x, \theta_B)$  satisfy a triangular array dominance condition.

**Assumption 11.**  $\nabla_{\theta_A}^2 \ln f_A(x, \theta_A)$  and  $\nabla_{\theta_B}^2 \ln f_B(x, \theta_B)$  satisfy a triangular array dominance condition.

**Assumption 12.**  $\ln f_k(x, \theta_k) \nabla_{\theta_l} \ln f_l(x, \theta_l)$  for  $k = A, B$  and  $l = A, B$  satisfy a triangular array dominance condition.

**Assumption 13.**  $E_{P_0}[\nabla_{\theta_k}^2 \ln f_k(X, \theta_k^*(P_0))]$  and  $E_{P_0}[\nabla_{\theta_k} \ln f_k(X_{0i}, \theta_k^*(P_0)) \nabla_{\theta_k} \ln f_k(X_{0i}, \theta_k^*(P_0))]$  for  $k = A, B$  are invertible.

**Assumption 14.** For some  $\delta > 0$ , we have  $\sup_{n \in \mathbb{N}} E_{P_n}[\|\nabla_{\theta_A} \ln f_A(X_{ni}, \theta_A^*(P_n))\|^{4+\delta}] < \infty$  and, similarly,  $\sup_{n \in \mathbb{N}} E_{P_n}[\|\nabla_{\theta_B} \ln f_B(X_{ni}, \theta_B^*(P_n))\|^{4+\delta}] < \infty$ .



**Assumption 15.** For some  $\delta > 0$ , we have  $\sup_{n \in \mathbb{N}} E_{P_n} [\|\ln f_A(X_{ni}, \theta_A^*(P_n))\|^{8+\delta}] < \infty$  and, similarly,  $\sup_{n \in \mathbb{N}} E_{P_n} [\|\ln f_B(X_{ni}, \theta_B^*(P_n))\|^{8+\delta}] < \infty$ .

**Assumption 16.** For some  $\delta > 0$ , we have  $\sup_{n \in \mathbb{N}} E_{P_n} [\|\nabla_{\theta_A}^2 \ln f_A(X_{ni}, \theta_A^*(P_n))\|^{4+\delta}] < \infty$  and, similarly,  $\sup_{n \in \mathbb{N}} E_{P_n} [\|\nabla_{\theta_B}^2 \ln f_B(X_{ni}, \theta_B^*(P_n))\|^{4+\delta}] < \infty$ .

**Assumption 17.**  $\nabla_{\theta_A}^3 \ln f_A(x, \theta_A)$  and  $\nabla_{\theta_B}^3 \ln f_A(x, \theta_B)$  satisfy a triangular array dominance condition.

**Assumption 18.**  $\sup_{n \in \mathbb{N}} E_{P_n} [\|\nabla_{\theta_k} \ln f_k(X_{ni}, \theta_k^*(P_n)) \nabla_{\theta_l} \ln f_l(X_{ni}, \theta_l^*(P_n))\|^{4+\delta}] < \infty$  for  $k = A, B$  and  $l = A, B$  for some  $\delta > 0$ .

**Assumption 19.**  $\nabla_k^2 \ln f_k(x, \theta_k) \nabla_{\theta_l} \ln f_l(x, \theta_l)$  for  $k = A, B$  and  $l = A, B$  satisfy a triangular array dominance condition.

**Theorem 5.** For  $\alpha \in (0, 1)$ , let  $z_{1-\alpha}$  denote the  $(1 - \alpha)$ -quantile of the standard normal distribution,  $\phi(\cdot)$  the standard normal density, and define  $M := \frac{1}{2} \phi(z_b)(z_b + z_{1-\alpha})c_\alpha$  with  $\Lambda := \max\{|\text{tr}(H_A^{-1}V_A)|, |\text{tr}(H_B^{-1}V_B)|\}$  and

$$c_\alpha := \sqrt{\frac{4\phi(z_{1-\alpha})\Lambda}{\phi(z_b)(z_b + z_{1-\alpha})(\sigma_A^2 + \sigma_B^2)}}.$$

Under Assumptions 1 and 9–19, if  $\hat{\varepsilon}_n$  is of the form (6), then, for any distribution  $P_0$  satisfying the null hypothesis, i.e.  $d^*(P_0) = 0$ ,

$$|P_0(\tilde{t}_n \geq z_{1-\alpha}) - \alpha| \leq Mn^{-1/4}\sqrt{\ln \ln n} + o\left(n^{-1/4}\sqrt{\ln \ln n}\right) \quad (7)$$

while, for sequences of local alternatives  $\{P_n\}$  satisfying  $P_0 = \lim_{n \rightarrow \infty} P_n$  and  $d^*(P_n) = \delta n^{-1/2}$  for any given  $\delta \in \mathbb{R}$ ,

$$P_n(t_n \geq z_{1-\alpha}) - P_n(\tilde{t}_n \geq z_{1-\alpha}) \leq Mn^{-1/4}\sqrt{\ln \ln n} + o\left(n^{-1/4}\sqrt{\ln \ln n}\right) \quad (8)$$

where  $t_n$  is the unregularized statistic as in (3). Moreover,  $\hat{\varepsilon}_n$  satisfies Assumption 7, and Assumption 8 with  $\mathcal{P}$  replaced by the set of distributions satisfying the assumptions of this theorem.

**Remark 6.** By Theorems 3 and 4 we expect  $\hat{\varepsilon}_n$  not to affect our test's first-order asymptotic properties such as power and size. Theorem 5 reflects this fact in the sense that both, size distortion (7) and power loss (8) converge to zero. Equation (6) provides the unique value of  $\hat{\varepsilon}_n$  that not only equalizes the rate at which both, size distortion and power loss, converge to zero, but also the constants in front of the rate.

**Remark 7.** Theorem 5 also verifies that the optimal epsilon (6) satisfies Assumptions 7 and 8, implying that all theorems in the previous sections hold with  $\hat{\varepsilon}_n$  replaced by the optimal expression in (6).

## 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 likelihood based on the KL criterion, i.i.d.

data and a t-statistic. In this section, we argue that our procedure applies much more generally and discuss some important, but mostly straightforward, extensions.

Our model selection test measures distance between the candidate models by KL distance. One could, however, consider other goodness-of-fit criteria such as in-sample or out-of-sample fit rather than KL distance. [Rivers and Vuong \(2002\)](#) propose such extensions of the Vuong test which would be completely analogous in our setting. An important example would be comparing the accuracy of competing forecasts. Consider two forecasts  $\{y_{(1)t}\}_{t=1}^T$  and  $\{y_{(2)t}\}_{t=1}^T$  of  $\{y_t\}_{t=1}^T$  and let  $\{e_{(k)t}\}_{t=1}^T$ ,  $k = 1, 2$ , be the corresponding forecast errors. In an influential paper, [Diebold and Mariano \(1995\)](#) discuss procedures for testing the hypothesis that the two forecasts are equally accurate, viz.

$$H_0 : Eg(e_{(1)t}) = Eg(e_{(2)t})$$

versus the alternative that the expectations are not equal, where  $g$  is some given loss function. [Diebold and Mariano \(1995\)](#) consider a test statistic  $\bar{d} := T^{-1/2} \sum_{t=1}^T [g(e_{(1)t}) - g(e_{(2)t})]$  which is asymptotically  $N(0, \sigma^2)$  under standard assumptions. Therefore, we can test  $H_0$  by simply comparing  $\bar{d}$  to a normal critical value. In this setting, we can apply our sample splitting scheme to obtain a test that is asymptotically uniformly of correct level, i.e. consider the modified statistic

$$\tilde{d} := \frac{T^{-1/2} \sum_{t=1}^T [\omega_t(\hat{\varepsilon}_T)g(e_{(1)t}) - \omega_{t+1}(\hat{\varepsilon}_T)g(e_{(2)t})]}{\sqrt{(1 + \hat{\varepsilon}_T)\hat{\sigma}^2 + \hat{\varepsilon}_T^2(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2}},$$

where  $\omega_t(\varepsilon)$  is defined as in (4), and  $\hat{\sigma}^2$ ,  $\hat{\sigma}_1^2$ , and  $\hat{\sigma}_2^2$  are estimators of  $\sigma^2$ , and the asymptotic variances of  $T^{-1/2} \sum_{t=1}^T g(e_{(1)t})$  and  $T^{-1/2} \sum_{t=1}^T g(e_{(2)t})$ , respectively.

A useful extension of theorems relaxes the i.i.d. assumption on the data generating process. In the case of comparing parametric likelihoods, our theory allows for conditional densities, so that time series dependence over a finite number of lags (e.g. AR(p)) can be accommodated simply by conditioning on the lagged variables. More generally, the limiting distribution of our test statistic ultimately only depends on the asymptotic normality of certain sample averages and it is clear that our results can easily be secured under a much wider range of conditions, including general stationary time series data.

Our testing procedure is based on estimating parameters from moment conditions. For simplicity of exposition we considered a Z-estimator which is simply the root of the empirical estimating equations. Clearly one could use any estimation procedure that estimates solutions to moment conditions. Our procedure requires only asymptotic normality of the resulting estimator which is readily established for a wide range of estimators (e.g. generalized method of moments (GMM), generalized empirical likelihood (GEL), minimum distance) using standard conditions available in the literature (see, for example, [Hansen \(1982\)](#), [Newey and McFadden \(1994\)](#), [Newey and Smith \(2004\)](#) and [van der Vaart \(1998\)](#)). Also, test statistics for testing  $H_0 : d^* = 0$  other than the t-statistic can be used, e.g. a Wald, Lagrange Multiplier or distance metric statistic. These are first-order asymptotically equivalent to our statistic under standard conditions.

In the present context, M-estimators are also attractive because terms can be added to the criterion function in order to penalize certain types of models. For example, one may want to avoid the selection of models with too many parameters and add a correction term that is increasing in the number of

parameters in a model. See, for instance, [Vuong \(1989, p. 318\)](#), [Sin and White \(1996\)](#) and references therein for correction terms that can be interpreted through information criteria such as AIC and BIC.

Interestingly, our method can also be extended to compare models defined by moment conditions rather than parametric likelihoods. In that case, one would replace the parametric scores  $E_{P_0}[\nabla_{\theta_A} \ln f_A(X; \theta_A^*)] = 0$  and  $E_{P_0}[\nabla_{\theta_B} \ln f_B(X; \theta_B^*)] = 0$  by the first-order derivatives of an empirical likelihood objective function and the KL-difference between the parametric densities by the difference in the respective objective functions. Other GEL objective functions could be used as well with the small difference being that they minimize divergence measures other than KL and so one may want to adjust our third moment condition accordingly. Notice, however, that comparisons based on GMM objective functions depend on the chosen weighting matrix and can, therefore, be very misleading ([Hall and Pelletier \(2011\)](#)).

We propose a regularization scheme which, in the observationally equivalent case, splits consecutive observations into two subsamples. The sample could, of course, be split in other ways as well. For example, one could consider the following reweighting scheme:

$$\hat{d} := \frac{1}{n} \sum_{i=1}^n \left( (1 + \varepsilon_{i,n}) \ln f_A(X_i; \hat{\theta}_A) - (1 - \varepsilon_{i,n}) \ln f_A(X_i; \hat{\theta}_B) \right)$$

where  $\varepsilon_{i,n}$  is an i.i.d. random variable independent of the sample and with a variance that shrinks to zero with the sample size  $n$ . This type of regularization does not assign special status to any observation, but on the other hand introduces more randomness, thereby reducing the power of the test. One could also deviate from our proposed even/odd splitting scheme and our procedure would work in the exact same way as discussed above. However, splitting into two halves is optimal in the sense that it minimizes the sum of the variances arising from the two half-samples. Furthermore, one can imagine splitting up the sample in many different ways and averaging over the resulting test statistics, but this procedure would lead to a complicated limiting distribution due to the nontrivial correlations among the individual statistics.

Finally, we could use our test to rank more than two models by incorporating it into a multiple testing framework in the usual way (see, for instance, the survey [Romano, Shaikh, and Wolf \(2010\)](#)).

## 8 Simulations

This section reports Monte Carlo simulation results for four pairs of models. The next section then shows how our procedure can be useful in an empirical application that has attracted a lot of attention in the past.

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 \(2013\)](#)'s modified Vuong test.<sup>9</sup> 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 (6). The two-step Vuong procedure for a level- $\alpha$  test is implemented by setting the level equal to  $\alpha$  in both individual steps.

---

<sup>9</sup>[Shi \(2013\)](#) 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 1** (Joint Normal Location). *This example is similar to one of Shi (2013)'s who constructed it in order to illustrate the potentially poor power of Vuong's test.*

$$P_0 := N\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

$$\mathcal{P}_A := \left\{ N\left(\begin{pmatrix} \mu_A \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) : \mu_A \in \Theta_A \right\}$$

$$\mathcal{P}_B := \left\{ N\left(\begin{pmatrix} 0 \\ \mu_B \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) : \mu_B \in \Theta_B \right\}$$

The null and alternative models are generated by varying  $\mu$  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.

**Example 2** (Misspecified Normals). *Let the true distribution of the random variables  $X_i$ ,  $i = 1, \dots, n$ , be  $N(\mu, 5)$ . The two parametric families to be compared are*

$$\mathcal{P}_A := \{N(\mu_A, 1) : \mu_A \in \Theta_A\}$$

$$\mathcal{P}_B := \{N(0, \sigma_B^2) : \sigma_B \in \Theta_B\}$$

The null and alternative models are generated by varying the true mean according to  $\mu = \sqrt{e^{2d+4} - 5}$  with  $d \in [-1, 1]$ . Both models are misspecified under the null ( $\mu_A^* = \sqrt{e^4 - 5}$  and  $\sigma_B^* = e^2$ ) and the alternatives. With  $\Theta_A$  not containing the origin, the two models are non-overlapping.

**Example 3** (Correctly Specified Normals). *Let the true distribution of the random variables  $X_i$ ,  $i = 1, \dots, n$ , be  $N(\mu, \sigma^2)$  and the two parametric families to be compared as in the previous example. The null and alternative models are generated by varying  $(\mu, \sigma^2)$  according to  $\mu = \sqrt{e^{2d-1+\sigma^2} - \sigma^2}$  with  $\sigma^2 \in [1, 5]$  and  $d \in [-1, 1]$ . The two models are correctly specified under the null ( $\mu_A = \mu = 0$ ,  $\sigma_B = \sigma = 1$ ), illustrating the case in which the two models overlap at the truth and thus are observationally equivalent under the null. Under the alternatives, they are both misspecified.*

**Example 4** (Nonnested Regressions). *This example is similar to one of Shi (2013)'s who constructed it in order to illustrate the potentially poor size control of Vuong's test. Let the random vector  $(Y_i, W_{i1}, \dots, W_{i10})$ ,  $i = 1, \dots, n$ , satisfy the regression equation*

$$Y_i = 1 + \frac{\tau}{\sqrt{9}} \sum_{k=1}^9 W_{ik} + \tau W_{i10} + \varepsilon_i, \quad \varepsilon_i \sim N(0, 2^2)$$

and  $(W_{i1}, \dots, W_{i14}) \sim N(0, I)$ . Consider model A,

$$Y_i = \alpha_0 + \sum_{k=1}^9 \alpha_k W_{ik} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_A^2),$$

and model B,

$$Y_i = \beta_0 + \beta_1 W_{i10} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_B^2).$$

For any value of  $\tau \neq 0$ , the two models have the same distance to the true model, but are both misspecified. We vary  $\tau$  in  $[0, 2]$ .

In Examples 1–3, we estimate means and variances with the sample means and variances and, in Example 4, we estimate the regressions by ordinary least-squares. Notice that these estimators are just the maximum-likelihood estimators in the particular models considered here. Table 1 reports the finite sample size of the different tests. Example 4 is the only one in which we consider a family of null hypotheses whereas, in all other examples, 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 2 shows the null rejection probabilities for Example 4 and Figures 3–6 the power curves for Examples 1–3. In all 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.

The two main findings from this simulation experiment can be summarized as follows. (i) In Table 1 and Figure 2, we see that all three tests control size well with our test having size very close to nominal size across all examples. Vuong’s and Shi’s test, on the other hand, tend to have size well below nominal size. (ii) Our new test and Shi’s test can have significantly higher power than Vuong’s test; see Figures 3 and 4. 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, our test can have higher (Figure 4) or lower (Figure 3) power than Shi’s test.

Besides the main findings we observe that our optimal epsilon choice better controls size and leads to a more powerful test compared to  $\varepsilon_n = 0.5$  and  $\varepsilon_n = 1$ , thus confirming the theoretical findings from Section 6. All tests perform well in the examples of misspecified and the correctly specified normals. They control size and all possess similar power curves.

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 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.<sup>10</sup> 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) and (10) with the following set of covariates:

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

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:

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

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.

Table 2 and the upper panel of Table 3 report the FIML estimates of  $\theta_1$  and  $\theta_2$ , respectively, with misspecification-robust (White (1982)) standard errors in parentheses. The lower panel of Table 3 shows the ML estimates of  $\theta_2$  based on equation (10) alone. The FIML estimates slightly differ from the two-step OLS estimates in Barro (1977). The differences arise from at least three different sources. First, joint estimation by FIML requires a slightly shorter sample for the money growth equation compared with Barro’s sample because data for the unemployment equation is available only for a shorter period. A lower  $R^2$  in the money growth equation is one noticeable consequence. More importantly, however, as in Barro’s results, the  $R^2$  values in the unemployment equation are relatively high, reflecting the consistency of the data with the NR/RE hypothesis. Second, FIML jointly estimates both equations and takes into account correlation among the residuals  $\varepsilon_{1t}$  and  $\varepsilon_{2t}$  whereas OLS does not. Third, numerical differences between the maximizer of the finite sample likelihood and the OLS estimator may occur.

<sup>10</sup>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.

The results of the pairwise model selection tests of new classical models versus Keynesian models are reported in Tables 4–11. Table 4 displays estimates of the optimal epsilon as defined in (6) and Table 5 the values of our test statistic  $\tilde{t}_n$  based on those estimated optimal epsilons. As a sensitivity analysis we also perform our test for epsilon values in a range from 0.1 to 1.2, which includes the optimal epsilon estimates and shows that the conclusions derived with the optimal epsilon do not change. Tables 7–9 report [Vuong \(1989\)](#)’s first-step statistic  $n\hat{\sigma}^2$ , the corresponding simulated critical values at 5% nominal level, and the second-step likelihood ratio statistic. Vuong’s two-step procedure rejects the null hypothesis of two models being equally far away from the truth when both of the two tests reject. Tables 10 and 11 show [Shi \(2013\)](#)’s test statistic and the corresponding critical values for 5% nominal level, respectively. When we compare Keynesian and new classical models based only the unemployment equation, all three tests fail to reject the 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. The Vuong test’s failure to distinguish the two theories is therefore less likely to be due to it underrejecting 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 do discriminate between the candidate new classical and Keynesian models. A larger sample or more 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, including or not including the money growth equation has no implications for whether the new classical or the Keynesian theory is superior to the other.

## 10 Conclusions

We propose a model selection test for choosing among parametric families of densities in the most general case in which no restrictions are imposed, neither on the relation between the two families nor on the relation between the families and the true distribution of the data. The test is based on a modified likelihood ratio statistic that is easy to compute and has an asymptotic standard normal distribution under all possible scenarios, regardless of whether none, one or both of the candidate models are misspecified, and regardless of whether they are overlapping, nested or nonnested. A consequence of this simple limit theory is that our test is able to uniformly controls size. We achieve this result by introducing a regularization parameter that ensures nondegeneracy of our test statistic in the limit. We provide a data-driven choice for this tuning parameter which optimally trades off power and size of our test. This optimality property translates into favorable finite sample properties as shown in our simulation study. In particular, the power gains relative to existing tests can be substantial. In the empirical section, we review the selection among Keynesian and new classical models of unemployment. Finally, we argue that the procedure can easily be extended to cover much more general types of models such as time series and moment condition models, for instance.



## A Simulations for Nested Models

In this section, we briefly demonstrate that our test also performs well for selecting among nested models. Typically, one can easily establish whether models are nested or not by inspection of the two parametric families. When they are in fact nested, the standard likelihood ratio test with a chi-square critical value is the most powerful test under well-known conditions.

**Example 5** (Nested Regressions with one Additional Regressor). *Let the random vector  $(Y_i, W_i, Z_i)$ ,  $i = 1, \dots, n$ , satisfy the regression equation*

$$Y_i = W_i + \tau W_i Z_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1)$$

*with  $W_i \sim N(3, 1)$ ,  $Z_i \sim N(0, 1)$  and  $\varepsilon_i \sim N(0, 1)$  all i.i.d. and mutually independent random variables. Consider model A,*

$$Y_i = \alpha_1 + \alpha_2 W_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_A^2),$$

*and model B,*

$$Y_i = \beta_1 + \beta_2 W_i + \beta_3 Z_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_B^2).$$

*Null and alternative models are generated by varying  $\tau$  over  $[0, 1.6]$ . Under the null ( $\tau = 0$ ), both models are correctly specified and model B nests model A while, under the alternatives, both are misspecified.*

**Example 6** (Nested Regressions with two Additional Regressors). *This example is similar to the previous one, except that model B has one more regressor, viz.*

$$Y_i = \beta_1 + \beta_2 W_i + \beta_3 Z_i + \beta_4 Z_i^2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_B^2),$$

*and the alternatives are generated from within model B:*

$$Y_i = W_i + \tau Z_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1).$$

*Therefore, the two models are nested, correctly specified under the null and the larger model is correctly specified even under the alternatives. This is the standard testing situation in which the second step of Vuong's procedure is equivalent to a Neyman Pearson ("NP") test of the hypothesis  $H_0 : \beta_3 = \beta_4 = 0$ .*

Figures 7 and 8 show the power plots for both examples. The lower two panels of Table 1 report the empirical rejection probabilities under the null. In both examples, compared to Vuong's and Shi's test, our test is more powerful for alternatives close to the null whereas the other two dominate for alternatives further away from the null. All three tests control size reasonably well, with Vuong's and Shi's test almost not rejecting under the null at all.

## B Proofs

For  $\theta = (\theta'_A, \theta'_B)'$ , let  $d_i(x; \theta, \varepsilon) := \omega_i(\varepsilon) \ln f_A(x; \theta_A) - \omega_{i+1}(\varepsilon) \ln f_B(x; \theta_B)$  and abbreviate  $d_i(\theta, \varepsilon) := d_i(X_{i,n}; \theta, \varepsilon)$ . Define  $\hat{G}(\theta) := \nabla_{\theta} \hat{g}(\theta)$  and  $G(\theta) := E_{P_0}[\nabla_{\theta} g(X; \theta)]$ .

**Lemma 1.** Suppose  $\{\varepsilon_n\} \in \mathcal{E}$ . Then, under any sequence  $P_n$  in  $\mathcal{P}$ ,

1.

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{d_i(\theta^*(P_n), \varepsilon_n) - (1 + \varepsilon_n/2)d^*(P_n)}{\tilde{\sigma}(P_n, \varepsilon_n)} \rightarrow_d N(0, 1).$$

2.

$$\frac{1}{n} \sum_{i=1}^n ((\ln f_k(X_{i,n}; \theta_k^*(P_n)))^2 - E_{P_n}[(\ln f_k(X_{i,n}; \theta_k^*(P_n)))^2]) = O_{P_n}(n^{-1/2}).$$

3.  $\hat{g}(\theta^*(P_n)) = O_{P_n}(n^{-1/2})$ .

*Proof.* For the first part, we start by showing that the following Lyapounov condition holds: for some  $\delta > 0$  as  $n \rightarrow \infty$ ,

$$\sum_{i=1}^{n/2} E_{P_n} \left[ \left| \frac{\Lambda_{2i-1}(P_n) + \Lambda_{2i}(P_n) + \varepsilon_n \Lambda_{2i,2i-1}(P_n) - (2 + \varepsilon_n)d^*(P_n)}{\sqrt{n}\tilde{\sigma}(P_n, \varepsilon_n)} \right|^{2+\delta} \right] \rightarrow 0, \quad (11)$$

where  $\Lambda_{i,j}(P) := \ln f_A(X_i; \theta^*(P)) - \ln f_B(X_j; \theta^*(P))$  and  $\Lambda_i(P) := \Lambda_{i,i}(P)$ . By the  $c_r$ -inequality,

$$\begin{aligned} & \sum_{i=1}^{n/2} E_{P_n} \left[ \left| \frac{\Lambda_{2i-1}(P_n) + \Lambda_{2i}(P_n) + \varepsilon_n \Lambda_{2i,2i-1}(P_n) - (2 + \varepsilon_n)d^*(P_n)}{\sqrt{n}\tilde{\sigma}(P_n, \varepsilon_n)} \right|^{2+\delta} \right] \\ & \leq \frac{2^{2+2\delta}}{n^{\delta/2}} \sum_{i=1}^{n/2} E_{P_n} \left[ |Z_{2i-1,n}|^{2+\delta} + |Z_{2i,n}|^{2+\delta} + |Z_{i,n,split}|^{2+\delta} + \left| \frac{(2 + \varepsilon_n)d^*(P_n)}{\tilde{\sigma}(P_n, \varepsilon_n)} \right|^{2+\delta} \right] \end{aligned} \quad (12)$$

with  $Z_{i,n} := \Lambda_i(P_n)/\tilde{\sigma}(P_n, \varepsilon_n)$  and  $Z_{i,n,split} := \varepsilon_n \Lambda_{2i,2i-1}(P_n)/\tilde{\sigma}(P_n, \varepsilon_n)$ . Consider the first of the four terms. If  $\sigma(P_n) \geq \underline{c}$  for some  $\underline{c} > 0$ , then

$$\begin{aligned} E_{P_n} \left[ |Z_{i,n}|^{2+\delta} \right] &= E_{P_n} \left[ \left| \frac{\ln f_A(X; \theta_A^*(P_n)) - \ln f_B(X; \theta_B^*(P_n))}{\tilde{\sigma}(P_n, \varepsilon_n)} \right|^{2+\delta} \right] \\ &\leq E_{P_n} \left[ \frac{|\ln f_A(X; \theta_A^*(P_n)) - \ln f_B(X; \theta_B^*(P_n))|^{2+\delta}}{(1 + \varepsilon_n)^{1+\delta/2} \sigma^{2+\delta}(P_n)} \right] \\ &\leq \frac{E_{P_n} [|D(X)|^{2+\delta}] \sigma^{2+\delta}(P_n)}{(1 + \varepsilon_n)^{1+\delta/2} \sigma^{2+\delta}(P_n)} \\ &= (1 + \varepsilon_n)^{-1-\delta/2} E_{P_n} [|D(X)|^{2+\delta}] \leq \overline{M} \end{aligned}$$

where the first inequality follows from the fact that  $\tilde{\sigma}^2(P, \varepsilon) = (1 + \varepsilon)\sigma^2(P) + \varepsilon^2(\sigma_A^2(P) + \sigma_B^2(P))/2$  is larger than either  $(1 + \varepsilon)\sigma^2(P)$  or  $\varepsilon^2(\sigma_A^2(P) + \sigma_B^2(P))/2$  (as  $\varepsilon \geq 0$ ). The second inequality is implied by the dominance condition (5). Since  $\overline{M}$  is independent of  $P_n$ , we have  $\sup_{n \geq 1} E_{P_n} [|Z_{2i-1,n}|^{2+\delta}] \leq \overline{M}$ , even if  $\sigma(P_n) \rightarrow 0$  as  $n \rightarrow \infty$ . Therefore, the first and second expectation in (12) are finite uniformly over  $n$ .

Next, consider the third expectation in (12):

$$\begin{aligned}
E_{P_n} \left[ |Z_{i,n,split}|^{2+\delta} \right] &= E_{P_n} \left[ \left| \frac{\varepsilon_n (\ln f_A(X_{2i}; \theta_A^*(P_n)) - \ln f_B(X_{2i-1}; \theta_B^*(P_n)))}{\tilde{\sigma}(P_n, \varepsilon_n)} \right|^{2+\delta} \right] \\
&\leq E_{P_n} \left[ \left| \frac{\varepsilon_n (\ln f_A(X_{2i}; \theta_A^*(P_n)) - \ln f_B(X_{2i-1}; \theta_B^*(P_n)))}{\varepsilon_n \sqrt{(\sigma_A^2(P_n) + \sigma_B^2(P_n))/2}} \right|^{2+\delta} \right] \\
&= E_{P_n} \left[ \left| \frac{\ln f_A(X_{2i}; \theta_A^*(P_n)) - \ln f_B(X_{2i-1}; \theta_B^*(P_n))}{\sqrt{(\sigma_A^2(P_n) + \sigma_B^2(P_n))/2}} \right|^{2+\delta} \right] \\
&\leq \underline{M}^{-1/2} 2^{1+\delta} \left\{ E_{P_n} \left[ |\ln f_A(X_{2i}; \theta_A^*(P_n))|^{2+\delta} \right] + E_{P_n} \left[ |\ln f_B(X_{2i-1}; \theta_B^*(P_n))|^{2+\delta} \right] \right\} \\
&\leq \underline{M}^{-1/2} 2^{2+\delta} \overline{M}
\end{aligned}$$

This bound is again valid uniformly over  $n$ .

Finally, by Lyapounov's Inequality, we have  $(1 + \varepsilon_n/2)d^*(P_n) \leq \tilde{\sigma}(P_n, \varepsilon_n)$ , uniformly in  $n$ , so that the fourth expectation in (12) is also finite, uniformly in  $n$ . In conclusion, we have established (11). Lyapounov's Central Limit Theorem (e.g. Theorem 23.11 in Davidson (1994)) then implies that, under any sequence  $P_n$  in  $\mathcal{P}$ ,

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{d_i(\theta^*(P_n), \varepsilon_n) - (1 + \varepsilon_n/2)d^*(P_n)}{\tilde{\sigma}(P_n, \varepsilon_n)} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n/2} \frac{\Lambda_{2i-1}(P_n) + \Lambda_{2i}(P_n) + \varepsilon_n \Lambda_{2i,2i-1}(P_n) - (2 + \varepsilon_n)d^*(P_n)}{\tilde{\sigma}(P_n, \varepsilon_n)} \rightarrow_d N(0, 1).
\end{aligned}$$

For the second part of the lemma, notice that

$$E_P \left[ \left| \frac{(\ln f_k(X; \theta_k^*(P)))^2 - E_P[(\ln f_k(X; \theta_k^*(P)))]^2}{\text{Var}_P((\ln f_k(X; \theta_k^*(P)))^2)^{1/2}} \right|^{2+\delta} \right] \leq \overline{M} \underline{M}^{-1}, \quad k = A, B, \quad (13)$$

for all  $P \in \mathcal{P}$  because  $\text{Var}_P((\ln f_k(X; \theta_k^*(P)))^2)$  is bounded away from zero by the definition of  $\mathcal{P}$  and because the numerator is bounded from above by  $\overline{M}$ . Therefore, we can apply the Lyapounov Central Limit Theorem as in the first part of the proof and the result follows. The third part of the lemma can be proved in exactly the same fashion as the second. Q.E.D.

**Lemma 2.** *Let  $X_{n,1}, \dots, X_{n,n}$  be an i.i.d. sample from  $P_n$  and Assumption 1 hold. Suppose there exists a unique  $\theta^*(P_n) \in \text{int}(\Theta)$  such that  $E_{P_n} g(X; \theta^*(P_n)) = 0$  and that there is a root  $\hat{\theta}$  satisfying the empirical analogue  $\hat{g}(\hat{\theta}) = 0$ . Further, assume the following conditions hold:*

(i)  $\hat{\varepsilon}_n$  is a sequence of measurable functions of  $X_{n,1}, \dots, X_{n,n}$  and there is a sequence  $\{\varepsilon_n\}$  in  $\mathcal{E}$  such that  $|\hat{\varepsilon}_n - \varepsilon_n| = o_{P_n}(1)$ .

(ii) For  $h(x; \theta)$  being any of the functions  $\ln f_k(x; \theta_k)$  and  $\nabla \ln f_k(x; \theta_k)$ ,  $k = A, B$ ,  $\theta = (\theta'_A, \theta'_B)'$ , we have

$$\sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n h(X_i; \theta) - E_{P_n} h(X_i; \theta) \right\| = o_{P_n}(1).$$

Then,  $\|\hat{\theta} - \theta^*(P_n)\| = o_{P_n}(1)$  and  $|\hat{d} - (1 + \varepsilon_n/2)d^*(P_n)| = o_{P_n}(1)$ .

*Proof.* Let  $\Psi_n(\theta) := E_{P_n}[g(X_{n,i}; \theta)]$ . The continuity of the moment function  $g(x; \cdot)$ , the compactness of  $\Theta$  and  $\theta^*(P_n)$  being the unique root of  $\Psi_n(\theta) = 0$  imply that, for any  $\kappa > 0$ ,

$$\inf_{\theta: \|\theta - \theta^*(P_n)\| \geq \kappa} \|\Psi_n(\theta)\| > 0.$$

The proof of  $\|\hat{\theta} - \theta^*(P_n)\| = o_{P_n}(1)$  then follows that of Theorem 5.9 in [van der Vaart \(1998\)](#). The second conclusion can be established as follows. A Taylor expansion around  $(\theta^*(P_n), \varepsilon_n)$  yields

$$\hat{d} = \frac{1}{n} \sum_{i=1}^n d_i(\hat{\theta}, \hat{\varepsilon}_n) = \frac{1}{n} \sum_{i=1}^n d_i(\theta^*(P_n), \varepsilon_n) + \frac{1}{n} \sum_{i=1}^n \nabla_{(\varepsilon, \theta)} d_i(\bar{\theta}_n, \bar{\varepsilon}_n) \begin{pmatrix} \hat{\varepsilon}_n - \varepsilon_n \\ \hat{\theta} - \theta^*(P_n) \end{pmatrix} = 0 \quad (14)$$

where  $(\bar{\theta}_n, \bar{\varepsilon}_n)$  lies on the line segment joining  $(\hat{\theta}, \hat{\varepsilon}_n)$  and  $(\theta^*(P_n), \varepsilon_n)$ . By (ii), the triangle inequality and  $\bar{\varepsilon}_n = O_{P_n}(1)$ , we have  $n^{-1} \sum_{i=1}^n \nabla_{(\varepsilon, \theta)} d_i(\bar{\theta}_n, \bar{\varepsilon}_n) = O_{P_n}(1)$ , so that

$$\left| \hat{d} - \frac{1}{n} \sum_{i=1}^n d_i(\theta^*(P_n), \varepsilon_n) \right| = o_{P_n}(1) \quad (15)$$

follows from  $\|\hat{\theta} - \theta^*(P_n)\| = o_{P_n}(1)$  and  $|\hat{\varepsilon}_n - \varepsilon_n| = o_{P_n}(1)$ . By (ii) and the triangle inequality, we also have

$$\left| \frac{1}{n} \sum_{i=1}^n d_i(\theta^*(P_n), \varepsilon_n) - \left(1 + \frac{\varepsilon_n}{2}\right) d^*(P_n) \right| = \left| \frac{1}{n} \sum_{i=1}^n d_i(\theta^*(P_n), \varepsilon_n) - E_{P_n} d_i(\theta^*(P_n), \varepsilon_n) \right| = o_{P_n}(1). \quad (16)$$

Together, (15) and (16) imply the second result. Q.E.D.

**Lemma 3.** *Let  $X_{n,1}, \dots, X_{n,n}$  be an i.i.d. sample from  $P_n$  and that the following conditions hold:*

(i)  $\hat{\varepsilon}_n$  is a sequence of measurable functions of  $X_{n,1}, \dots, X_{n,n}$  such that there is a sequence  $\{\varepsilon_n\}$  in  $\mathcal{E}$  satisfying  $|\hat{\varepsilon}_n - \varepsilon_n| = O_{P_n}(n^{-1/2})$ .

(ii) For  $h(x; \theta)$  being any of the functions  $\ln f_k(X; \theta_k)$ ,  $\ln f_k(x; \theta_k) \nabla \ln f_j(x; \theta_j)$ , and  $\nabla \ln f_k(X; \theta_k)$ ,  $j, k = A, B$ ,  $\theta = (\theta'_A, \theta'_B)'$ , we have

$$\sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n h(X_i; \theta) - E_{P_n} h(X_i; \theta) \right\| = o_{P_n}(1),$$

and

$$\frac{1}{n} \sum_{i=1}^n ((\ln f_k(X_{i,n}; \theta_k^*(P_n)))^2 - E_{P_n}[(\ln f_k(X_{i,n}; \theta_k^*(P_n)))^2]) = O_{P_n}(n^{-1/2}).$$

(iii)  $\|\hat{\theta} - \theta^*(P_n)\| = O_{P_n}(n^{-1/2})$ ,

(iv) There are constants  $0 < \underline{M} \leq \overline{M} < \infty$  such that  $\underline{M} \leq \sigma_k(P_n) \leq \overline{M}$  for all  $n$  and  $k = A, B$ .

Then, for  $\hat{\sigma}^2 = \hat{\sigma}^2(\hat{\theta}, \hat{\varepsilon}_n)$ ,

$$\left| \frac{\tilde{\sigma}^2(P_n, \varepsilon_n)}{\hat{\sigma}^2} - 1 \right| \rightarrow_{P_n} 0.$$

*Proof.* First, we establish

$$|\hat{\sigma}^2 - \sigma^2(P_n)| = O_{P_n}(n^{-1/2}) \quad \text{and} \quad |\hat{\sigma}_k^2 - \sigma_k^2(P_n)| = O_{P_n}(n^{-1/2}), k = A, B. \quad (17)$$

Notice that by a Taylor expansion around  $\theta^*(P_n)$ , under  $P_n$ , we have

$$|\hat{\sigma}^2 - \hat{\sigma}^2(\theta^*(P_n))| \leq \left| \nabla_{\theta} \hat{\sigma}^2(\bar{\theta}_n) \left( \hat{\theta} - \theta^*(P_n) \right) \right| = O_{P_n}(n^{-1/2})$$

where  $\bar{\theta}_n$  lies on the line segment joining  $\hat{\theta}$  and  $\theta^*(P_n)$ . Uniform convergence of  $\ln f_k(X; \theta_k)$ ,  $\nabla \ln f_k(X; \theta_k)$  and  $\ln f_k(x; \theta_k) \nabla \ln f_j(x; \theta_j)$ ,  $j, k = A, B$ , in (ii) together with the Cauchy-Schwartz inequality imply  $\|\nabla_{\theta} \hat{\sigma}^2(\bar{\theta}_n)\| = O_{P_n}(1)$  so that the equality above follows from the consistency requirement in (iii). Similarly,  $|\hat{\sigma}_k^2 - \hat{\sigma}_k^2(\theta_k^*(P_n))| = O_{P_n}(n^{-1/2})$  for  $k = A, B$ . By the second part of (ii) and the Hölder inequality,  $|\hat{\sigma}_k^2(\theta^*(P_n)) - \sigma_k^2(P_n)| = O_{P_n}(n^{-1/2})$  for  $k = A, B$ , and the desired result (17) follows.

The remainder of the proof separately treats the two cases  $\sigma^2(P_n) \rightarrow \sigma_{\infty}^2 > 0$  and  $\sigma^2(P_n) \rightarrow 0$ . First, consider  $\sigma^2(P_n) \rightarrow \sigma_{\infty}^2 > 0$ . In this case, by (iv) and the definition of  $\mathcal{E}$ ,  $\tilde{\sigma}^2(P_n, \varepsilon_n)$  also converges to a finite, nonzero constant. Thus, (17) and (i) directly yield  $|\hat{\sigma}^2 - \tilde{\sigma}^2(P_n, \varepsilon_n)| = O_{P_n}(n^{-1/2})$  so that

$$\left| \frac{\tilde{\sigma}^2(P_n, \varepsilon_n)}{\hat{\sigma}^2} - 1 \right| = \left| \frac{\tilde{\sigma}^2(P_n, \varepsilon_n)}{\tilde{\sigma}^2(P_n, \varepsilon_n) + O_{P_n}(n^{-1/2})} - 1 \right| = o_{P_n}(1).$$

Now, consider  $\sigma^2(P_n) \rightarrow 0$ . We further split this case into three subcases: (a)  $\sigma^2(P_n)/\varepsilon_n^2 \rightarrow 0$  which means that either  $\sigma^2(P_n)$  and  $\varepsilon_n^2$  both vanish, but  $\sigma^2(P_n)$  at a faster rate, or that  $\sigma^2(P_n)$  converges to zero at an arbitrary rate while  $\varepsilon_n^2$  stays bounded away from zero; (b)  $\sigma^2(P_n)/\varepsilon_n^2 \rightarrow \infty$ , i.e.  $\sigma^2(P_n)$  and  $\varepsilon_n^2$  both vanish, but  $\varepsilon_n^2$  at a faster rate; (c)  $\sigma^2(P_n)/\varepsilon_n^2 \rightarrow c \neq 0$ , i.e. both vanish at the same rate.

Consider subcase (a). By Assumption (i), we have

$$\frac{\hat{\varepsilon}_n}{\varepsilon_n} = 1 + \frac{\hat{\varepsilon}_n - \varepsilon_n}{\varepsilon_n} = 1 + O_{P_n}(n^{-1/2}\varepsilon_n^{-1}) = 1 + o_{P_n}(1).$$

Similarly, by (17),

$$\frac{\hat{\sigma}^2}{\varepsilon_n^2} = \frac{\sigma^2(P_n)}{\varepsilon_n^2} + \frac{\hat{\sigma}^2 - \sigma^2(P_n)}{\varepsilon_n^2} = o(1) + O_{P_n}(n^{-1/2}\varepsilon_n^{-2}) = o_{P_n}(1).$$

Therefore,

$$\begin{aligned} \frac{\tilde{\sigma}^2(P_n, \varepsilon_n)}{\hat{\sigma}^2} &= \frac{(1 + \varepsilon_n)\sigma^2(P_n) + \frac{\varepsilon_n^2}{2}(\sigma_A^2(P_n) + \sigma_B^2(P_n))}{(1 + \hat{\varepsilon}_n)\hat{\sigma}^2 + \frac{\hat{\varepsilon}_n^2}{2}(\hat{\sigma}_A^2 + \hat{\sigma}_B^2)} = \frac{\frac{1}{2}(\sigma_A^2(P_n) + \sigma_B^2(P_n)) + O(\frac{\sigma^2(P_n)}{\varepsilon_n^2})}{\frac{\hat{\varepsilon}_n^2}{\varepsilon_n^2}(\hat{\sigma}_A^2 + \hat{\sigma}_B^2) + O_{P_n}(\frac{\hat{\sigma}^2}{\varepsilon_n^2})} \\ &= \frac{\frac{1}{2}(\sigma_A^2(P_n) + \sigma_B^2(P_n)) + o(1)}{\frac{1}{2}(\sigma_A^2(P_n) + \sigma_B^2(P_n)) + o_{P_n}(1)} = 1 + o_{P_n}(1) \end{aligned}$$

In subcase (b), we use a similar reasoning as above to show that  $\hat{\varepsilon}_n^2/\sigma^2(P_n) = o_{P_n}(1)$  and  $\hat{\sigma}^2/\sigma^2(P_n) = 1 + o_{P_n}(1)$ . Therefore,

$$\begin{aligned} \frac{\tilde{\sigma}^2(P_n, \varepsilon_n)}{\hat{\sigma}^2} &= \frac{(1 + \varepsilon_n)\sigma^2(P_n) + \frac{\varepsilon_n^2}{2}(\sigma_A^2(P_n) + \sigma_B^2(P_n))}{(1 + \hat{\varepsilon}_n)\hat{\sigma}^2 + \frac{\hat{\varepsilon}_n^2}{2}(\hat{\sigma}_A^2 + \hat{\sigma}_B^2)} = \frac{(1 + \varepsilon_n) + O(\varepsilon_n^2/\sigma^2(P_n))}{(1 + \hat{\varepsilon}_n)\frac{\hat{\sigma}^2}{\sigma^2(P_n)} + O_{P_n}(\hat{\varepsilon}_n^2/\sigma^2(P_n))} \\ &= \frac{1 + o(1)}{1 + o_{P_n}(1)} = 1 + o_{P_n}(1) \end{aligned}$$

In subcase (c), we also have  $\hat{\sigma}^2/\sigma^2(P_n) = 1 + o_{P_n}(1)$  and  $\hat{\varepsilon}_n^2/\sigma^2(P_n) = \varepsilon_n^2/\sigma^2(P_n) + o_{P_n}(1)$  so that

$$\begin{aligned} \frac{\tilde{\sigma}^2(P_n, \varepsilon_n)}{\hat{\sigma}^2} &= \frac{(1 + \varepsilon_n)\sigma^2(P_n) + \frac{\varepsilon_n^2}{2}(\sigma_A^2(P_n) + \sigma_B^2(P_n))}{(1 + \hat{\varepsilon}_n)\hat{\sigma}^2 + \frac{\hat{\varepsilon}_n^2}{2}(\hat{\sigma}_A^2 + \hat{\sigma}_B^2)} \\ &= \frac{\sigma^2(P_n) + \frac{\varepsilon_n^2}{2}(\sigma_A^2(P_n) + \sigma_B^2(P_n)) + o(\varepsilon_n^2)}{\hat{\sigma}^2 + \frac{\hat{\varepsilon}_n^2}{2}(\hat{\sigma}_A^2 + \hat{\sigma}_B^2) + O_{P_n}(\hat{\varepsilon}_n\hat{\sigma}^2)} \\ &= \frac{1 + \frac{\varepsilon_n^2}{2\sigma^2(P_n)}(\sigma_A^2(P_n) + \sigma_B^2(P_n)) + o(\varepsilon_n^2/\sigma^2(P_n))}{\frac{\hat{\sigma}^2}{\sigma^2(P_n)} + \frac{\hat{\varepsilon}_n^2}{2\sigma^2(P_n)}(\sigma_A^2(P_n) + \sigma_B^2(P_n)) + o_{P_n}(\varepsilon_n^2/\sigma^2(P_n))} \\ &= 1 + o_{P_n}(1) \end{aligned}$$

which uses the fact that  $O_{P_n}(\hat{\varepsilon}_n\hat{\sigma}^2) = O_{P_n}(\varepsilon_n(\sigma^2(P_n) + n^{-1/2})) = o_{P_n}(\varepsilon_n^2)$ . Q.E.D.

**Lemma 4.** *Suppose Assumption 1 holds. Let  $\hat{\varepsilon}_n$  be a sequence of real-valued, measurable functions of the triangular array  $X_{n,1}, \dots, X_{n,n}$ , an i.i.d. sample from  $P_n$ , and  $\mathcal{Q}$  be some subset of  $\mathcal{P}$ . Assume that, for every sequence  $\{P_n\}$  in  $\mathcal{Q}$ , there is a sequence  $\{\varepsilon_n\} \in \mathcal{E}$  with  $|\hat{\varepsilon}_n - \varepsilon_n| = O_{P_n}(n^{-1/2})$ . Let  $\bar{\delta} \in [-\infty, +\infty]$  be such that  $\sqrt{n}d^*(P_n)(1 + \varepsilon_n/2)/\tilde{\sigma}(P_n, \varepsilon_n) \rightarrow \bar{\delta}$ . Then, under any sequence  $\{P_n\}$  in  $\mathcal{Q}$ , if  $|\bar{\delta}| < \infty$ ,*

$$\frac{\sqrt{n}\hat{d}}{\hat{\sigma}} \rightarrow_d N(\bar{\delta}, 1).$$

If  $|\bar{\delta}| = \infty$ , then  $|\sqrt{n}\hat{d}/\hat{\sigma}| \rightarrow_{P_n} \infty$ .

*Proof.* Suppose  $|\bar{\delta}| < \infty$ . First, we establish two auxiliary results, viz. the orders of  $\tilde{\sigma}(P_n, \varepsilon_n)^{-1}$  and  $\hat{\theta} - \theta^*(P_n)$ . To that end, consider two cases: (a)  $P_n$  approaches the observationally equivalent case, i.e.  $\sigma(P_n) \rightarrow 0$ ; (b)  $P_n$  satisfies  $\sigma(P_n) \rightarrow c \neq 0$ . In the first case, since by part (iv) of Definition 2,  $\sigma_k^2(P_n)$  is bounded away from zero and  $n^{1/4}\varepsilon_n \rightarrow \infty$ ,

$$n\tilde{\sigma}^2(P_n, \varepsilon_n) = n(1 + \varepsilon_n)\sigma^2(P_n) + n\varepsilon_n^2(\sigma_A^2(P_n) + \sigma_B^2(P_n))/2 \rightarrow \infty$$

so that  $\tilde{\sigma}(P_n, \varepsilon_n)^{-1} = o(n^{1/2})$ . In the second case,  $\tilde{\sigma}(P_n, \varepsilon_n) \rightarrow c \neq 0$  so that  $\tilde{\sigma}(P_n, \varepsilon_n)^{-1} = O(1) = o(n^{1/2})$ . In conclusion,

$$\tilde{\sigma}(P_n, \varepsilon_n)^{-1} = o(n^{1/2}). \tag{18}$$

Next, consider the order of  $\hat{\theta} - \theta^*(P_n)$ . A Taylor expansion with  $\bar{\theta}$  on the line segment joining  $\hat{\theta}$  and  $\theta^*(P_n)$  yields  $\hat{\theta} - \theta^*(P_n) = -\hat{G}(\bar{\theta})^{-1}\hat{g}(\theta^*(P_n))$ . By Assumption 1, parts (i) and (iii) of Definition 2, and

Lemma 2.4 of [Newey and McFadden \(1994\)](#),  $\hat{G}(\theta)$  converges in probability, under  $P_n$ , uniformly over  $\Theta$ . Part (i) and (iii) of Definition 2 together with Assumption 1 imply Assumption (ii) of Lemma 2, so that we can use it to obtain consistency of  $\hat{\theta}$  and  $\bar{\theta}$  under  $P_n$ . Therefore, letting  $G_P(\theta) := E_P[\nabla_{\theta}g(X; \theta)]$  and  $G_n := G_{P_n}(\theta^*(P_n))$ , we have

$$\begin{aligned} \left\| \hat{G}(\bar{\theta}) - G_n \right\| &\leq \left\| \hat{G}(\bar{\theta}) - G_{P_n}(\bar{\theta}) \right\| + \left\| G_{P_n}(\bar{\theta}) - G_n \right\| \\ &\leq \sup_{\theta \in \Theta} \left\| \hat{G}(\theta) - G_{P_n}(\theta) \right\| + o_{P_n}(1) = o_{P_n}(1). \end{aligned}$$

Furthermore, by (iv) of Definition 2,  $\hat{G}(\bar{\theta})$  is invertible with probability approaching one, under  $P_n$ . By part 3. of Lemma 1,  $\hat{g}(\theta^*(P_n)) = O_{P_n}(n^{-1/2})$ , so that, in conclusion,

$$\hat{\theta} - \theta^*(P_n) = -\hat{G}(\bar{\theta})^{-1} \hat{g}(\theta^*(P_n)) = O_{P_n}(n^{-1/2}). \quad (19)$$

With the auxiliary results established, we now consider the following decomposition:

$$\frac{\sqrt{n}\hat{d}}{\tilde{\sigma}(P_n, \varepsilon_n)} = \frac{\sqrt{n}d^*(P_n)(1 + \frac{\varepsilon_n}{2})}{\tilde{\sigma}(P_n, \varepsilon_n)} + \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \left( d_i(\hat{\theta}, \hat{\varepsilon}_n) - d^*(P_n)(1 + \frac{\varepsilon_n}{2}) \right)}{\tilde{\sigma}(P_n, \varepsilon_n)}.$$

The assumption  $|\hat{\varepsilon}_n - \varepsilon_n| = O_{P_n}(n^{-1/2})$  and a Taylor expansion of  $d_i(\hat{\theta}, \hat{\varepsilon}_n) - d^*(P_n)(1 + \hat{\varepsilon}_n/2)$  around  $(\theta^*(P_n), \varepsilon_n)$  yield

$$\begin{aligned} \frac{\sqrt{n}\hat{d}}{\tilde{\sigma}(P_n, \varepsilon_n)} &= \bar{\delta} + o_{P_n}(1) + \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \left( d_i(\theta^*(P_n), \varepsilon_n) - d^*(P_n)(1 + \frac{\varepsilon_n}{2}) \right)}{\tilde{\sigma}(P_n, \varepsilon_n)} \\ &\quad + \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_{\theta} d_i(\theta^*(P_n), \varepsilon_n) (\hat{\theta} - \theta^*(P_n))}{\tilde{\sigma}(P_n, \varepsilon_n)} \\ &\quad + \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \nabla_{\varepsilon} d_i(\theta^*(P_n), \varepsilon_n) - \frac{1}{2} d^*(P_n) \right) (\hat{\varepsilon}_n - \varepsilon_n)}{\tilde{\sigma}(P_n, \varepsilon_n)} + R_n \end{aligned} \quad (20)$$

where, for some  $(\bar{\theta}_n, \bar{\varepsilon}_n)$  on the line segment joining  $(\hat{\theta}, \hat{\varepsilon}_n)$  and  $(\theta^*(P_n), \varepsilon_n)$ ,

$$\begin{aligned} |R_n| &\leq \sqrt{n} \tilde{\sigma}(P_n, \varepsilon_n)^{-1} \left\| \frac{1}{n} \sum_{i=1}^n \nabla_{\hat{\theta}}^2 d_i(\bar{\theta}_n, \bar{\varepsilon}_n) \right\| \left\| \hat{\theta} - \theta^*(P_n) \right\|^2 \\ &\quad + \sqrt{n} \tilde{\sigma}(P_n, \varepsilon_n)^{-1} \left\| \frac{1}{n} \sum_{i=1}^n \nabla_{\hat{\varepsilon}}^2 d_i(\bar{\theta}_n, \bar{\varepsilon}_n) \right\| \left\| \hat{\varepsilon}_n - \varepsilon_n \right\|^2 \\ &= \sqrt{n} o(n^{1/2}) O_{P_n}(1) O_{P_n}(n^{-1}) + 0 = o_{P_n}(1). \end{aligned}$$

The first equality holds for the following reason. By Assumption 1, parts (i) and (iii) of Definition 2, and Lemma 2.4 of [Newey and McFadden \(1994\)](#),  $\|n^{-1} \sum_{i=1}^n \nabla_{\theta}^2 \ln f_k(X_{n,i}; \theta)\|$ ,  $k = A, B$ , converges in probability, under  $P_n$ , uniformly over  $\Theta$ . By the triangle inequality and the fact that  $\hat{\varepsilon}_n = O_{P_n}(1)$ , and thus  $\bar{\varepsilon}_n = O_{P_n}(1)$ , we also have  $\|n^{-1} \sum_{i=1}^n \nabla_{\hat{\theta}}^2 d_i(\bar{\theta}_n, \bar{\varepsilon}_n)\| = O_{P_n}(1)$ . (18), (19), and the assumption  $|\hat{\varepsilon}_n - \varepsilon_n| = O_{P_n}(n^{-1/2})$  then imply the equality.

We now separately consider each of the remaining three terms in (20). By part 1. of Lemma 1, the first term is asymptotically  $N(0, 1)$  under  $P_n$ . For the second term, notice that  $n^{-1} \sum_{i=1}^n \nabla_{\theta} d_i(\theta^*(P_n), \varepsilon_n)$  is a linear transformation of  $\hat{g}(\theta^*(P_n))$  and, thus by part 3. of Lemma 1 and  $\varepsilon_n = O(1)$ ,  $O_{P_n}(n^{-1/2})$ . Therefore, (18) and (19) imply

$$\frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_{\theta} d_i(\theta^*(P_n), \varepsilon_n)(\hat{\theta} - \theta^*(P_n))}{\tilde{\sigma}(P_n, \varepsilon_n)} = O_{P_n}(1)O_{P_n}(n^{-1/2})o(n^{1/2}) = o_{P_n}(1).$$

In the third term,

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left( \nabla_{\varepsilon} d_i(\theta^*(P_n), \varepsilon_n) - \frac{d^*(P_n)}{2} \right) \\ &= \frac{1}{n} \sum_{i=1}^n \left( \ln f_A(X_{n,2i-1}; \theta_A^*(P_n)) - \ln f_B(X_{n,2i}; \theta_B^*(P_n)) - \frac{d^*(P_n)}{2} \right) = O_{P_n}(n^{-1/2}) \end{aligned}$$

by a similar argument as in part 1. of Lemma 1, so that

$$\frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n (\nabla_{\varepsilon} d_i(\theta^*(P_n), \varepsilon_n) - \frac{1}{2} d^*(P_n)) (\hat{\varepsilon}_n - \varepsilon_n)}{\tilde{\sigma}(P_n, \varepsilon_n)} = O_{P_n}(1)O_{P_n}(n^{-1/2})o(n^{1/2}) = o_{P_n}(1).$$

In conclusion,  $\sqrt{n}\hat{d}/\tilde{\sigma}(P_n, \varepsilon_n) \rightarrow_d N(\bar{\delta}, 1)$  under  $P_n$ . The corresponding result with the estimated standard deviation,  $\hat{\sigma}$ , in the denominator rather than  $\tilde{\sigma}(P_n, \varepsilon_n)$  follows from Lemma 3, using (19). The case  $|\bar{\delta}| = \infty$  follows from a similar argument. Q.E.D.

*Proof of Theorem 1.* Assumptions 1, 2(i) and 3(ii.a) imply the conditions of Lemma 2.4 in Newey and McFadden (1994) so that  $n^{-1} \sum_{i=1}^n \ln f_k(x; \theta_k)$  converges uniformly in probability over  $\Theta_k$ . By Assumptions 1 and 2(ii), for any  $\kappa > 0$ ,  $\inf_{\theta: \|\theta - \theta^*\| \geq \kappa} \|E_{P_0} g(X; \theta)\| > 0$  so that we can apply Theorem 5.9 in van der Vaart (1998) to show consistency of  $\hat{\theta}$ .

The standard Taylor expansion argument with  $\bar{\theta}$  on the line segment joining  $\hat{\theta}$  and  $\theta^*$  yields  $\sqrt{n}(\hat{\theta} - \theta^*) = \hat{G}(\bar{\theta})^{-1} \sqrt{n}\hat{g}(\theta^*)$ . Assumptions 1, 2(i) and 3(ii.c) imply the conditions of Lemma 2.4 in Newey and McFadden (1994) so that  $n^{-1} \sum_{i=1}^n \nabla_{\theta_k}^2 \ln f_k(x; \theta_k)$  converges uniformly in probability over  $\Theta_k$ . Since  $\hat{\theta}$  is consistent,  $\bar{\theta}$  is as well and  $\hat{G}(\bar{\theta}) \rightarrow_p G(\theta^*)$ . The limit is invertible by Assumption 2(iii) and, thus,  $\hat{G}(\bar{\theta})$  is invertible with probability approaching one. Furthermore, by the finite variance in Assumption 3(i), the CLT implies that  $\sqrt{n}\hat{g}(\theta^*)$  is asymptotically normal, so that  $\sqrt{n}(\hat{\theta} - \theta^*) = O_{P_0}(1)$ .

Assumptions 1, 2(i) and 3(ii.b) imply the conditions of Lemma 2.4 in Newey and McFadden (1994) so that  $n^{-1} \sum_{i=1}^n \nabla_{\theta_k} \ln f_k(x; \theta_k)$  converges uniformly in probability over  $\Theta_k$ . Furthermore, by the finite variance in Assumption 3(i), the CLT implies that  $n^{-1/2} \sum_{i=1}^n d_i(\theta^*)$  is asymptotically normal, so that

$$\sqrt{n}\hat{d} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[ d_i(\theta^*) + \nabla_{\theta} d_i(\bar{\theta}) (\hat{\theta} - \theta^*) \right] = \frac{1}{\sqrt{n}} \sum_{i=1}^n d_i(\theta^*) + o_{P_0}(1) \rightarrow_d N(0, \sigma^2).$$

By Assumptions 1, 2(i), 3(i), consistency of  $\hat{\theta}$ , and Lemma 4.3 of Newey and McFadden (1994),  $\hat{\sigma}_k^2$ ,  $k = A, B$ , and  $\hat{\sigma}_{AB}$  are consistent estimators, so that  $\hat{\sigma}$  is consistent as well. Slutsky's Theorem then yields the desired result. Q.E.D.



*Proof of Theorem 2.* We show the result by applying Lemma 4. Let  $\mathcal{Q} = \{P_0\}$  and  $\bar{\delta} = 0$ . Part (i) of Definition 2 holds by Assumption 2(ii). Part (iii) by Assumption 6(ii). Finally, part (iv) of Definition 2 holds because of Assumptions 2(iii) and 5, and Assumption 7 implies Assumption 8. The uniform moment bounds in (ii) of Definition 2 hold because of Assumption 6(i).

It remains to show that the dominance condition (5) in (ii) of Definition 2 holds. This can be seen as follows. In the non-overlapping case,  $\sigma^2 > 0$ , (5) is implied by Assumption 6(i). In the overlapping case, the information matrix equality holds, so that  $\text{Var}_{P_0}(\nabla_{\theta_k} \ln f_k(X; \theta_k^*)) = E_{P_0}[\nabla_{\theta_k}^2 \ln f_k(X; \theta_k^*)]$ ,  $k = A, B$ , is invertible by Assumption 2(iii). Let  $\lambda_{\min}$  be the minimum of the eigenvalues of both matrices and note that it must be strictly larger than zero. Then it is easy to show that (5) holds for  $D(x) := \sqrt{2}\bar{F}_2(x)/\lambda_{\min}$  because of Assumption 6(iii). Q.E.D.

*Proof of Theorem 3.* Lemma 4, whose assumptions are satisfied by setting  $\mathcal{Q} = \mathcal{P}_0$  and by Assumptions 1 and 8, implies that  $\sqrt{nd}/\hat{\sigma} \rightarrow_d N(\bar{\delta}, 1)$  under any sequence  $\{P_n\}$  in  $\mathcal{P}_0$ . Using this result, the theorem follows from analogous reasoning as in the proof of Theorem 11.4.5 of Lehmann and Romano (2005). Q.E.D.

*Proof of Theorem 4.* The result follows directly from Lemma 4. Q.E.D.

*Proof of Theorem 5.* The proof proceeds by decomposing the statistic into an asymptotically normal component and non-normal remainder terms that are negligible in an almost sure sense. We first obtain some generic asymptotic expansions that hold for triangular arrays (as needed for local power calculation). These expansions, specialized to the case of sequences, are also used for size calculations.

We first observe that, by Assumptions 9 and 10, Lemma 5 implies that  $n^{-1} \sum_{i=1}^n \ln f_A(X_{ni}, \theta_A)$  converges almost surely uniformly for all  $\theta_A \in \Theta_A$  to  $E_{P_0}[\ln f_A(X_{0i}, \theta_A)]$ . This in turn implies that  $\hat{\theta}_A \rightarrow_{as} \theta_A^* := \theta_A^*(P_0)$  by the usual argument for consistency of MLE, adapted for almost sure convergence. We then expand the first order condition for  $\hat{\theta}_A$  as

$$0 = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta_A} \ln f_A(X_{ni}, \hat{\theta}_A) = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta_A} \ln f_A(X_{ni}, \theta_A^*) + \frac{1}{n} \sum_{i=1}^n \nabla_{\theta_A}^2 \ln f_A(X_{ni}, \bar{\theta}_A) (\hat{\theta}_A - \theta_A^*)$$

where  $\bar{\theta}_A$  is a mean value on the line segment joining  $\hat{\theta}_A$  and  $\theta_A^*$ . By Assumptions 9 and 11, Lemma 5 implies that  $n^{-1} \sum_{i=1}^n \nabla_{\theta_A}^2 \ln f_A(X_{ni}, \theta_A)$  converges uniformly to  $E_{P_0}[\nabla_{\theta_A}^2 \ln f_A(X_{0i}, \theta_A)]$  for all  $\theta_A \in \Theta_A$ . Since  $n^{-1} \sum_{i=1}^n \nabla_{\theta_A}^2 \ln f_A(X_{ni}, \theta_A)$  is continuous in  $\theta_A$  at each  $n$  by Assumption 11 and the convergence is uniform, it follows that the limit  $E_{P_0}[\nabla_{\theta_A}^2 \ln f_A(X_{0i}, \theta_A)]$  is also continuous in  $\theta_A$ . Since  $\hat{\theta}_A \rightarrow_{as} \theta_A^*$  and therefore  $\bar{\theta}_A \rightarrow_{as} \theta_A^*$ , we have

$$\frac{1}{n} \sum_{i=1}^n \nabla_{\theta_A}^2 \ln f_A(X_{ni}, \bar{\theta}_A) = E_{P_0}[\nabla_{\theta_A}^2 \ln f_A(X_{0i}, \theta_A^*)] + o_{as}(1)$$

and it follows, under Assumption 13, that

$$\hat{\theta}_A - \theta_A^* = - \left( (E_{P_0}[\nabla_{\theta_A}^2 \ln f_A(X_{0i}, \theta_A^*)])^{-1} + o_{as}(1) \right) \frac{1}{n} \sum_{i=1}^n \nabla_{\theta_A} \ln f_A(X_{ni}, \theta_A^*). \quad (21)$$

Let  $\|M\|_F$  denote the largest eigenvalue of matrix  $M$ . Observe that, by Assumption 18 and dominated convergence,  $V_A(P_n) \rightarrow V_A$  with  $\|V_A\|_F < \infty$ . Moreover  $V_A$  is invertible by Assumption 13. We can then write

$$\begin{aligned} \limsup_{n \rightarrow \infty} \left| \frac{1}{n} \sum_{i=1}^n \nabla_{\theta_{A_j}} \ln f_A(X_{ni}, \theta_A^*) \right| &= \limsup_{n \rightarrow \infty} \left| V_A(P_n)^{1/2} V_A(P_n)^{-1/2} \frac{1}{n} \sum_{i=1}^n \nabla_{\theta_{A_j}} \ln f_A(X_{ni}, \theta_A^*) \right| \\ &= \left( \lim_{n \rightarrow \infty} V_A(P_n)^{1/2} \right) \left( \limsup_{n \rightarrow \infty} \left| V_A(P_n)^{-1/2} \frac{1}{n} \sum_{i=1}^n \nabla_{\theta_{A_j}} \ln f_A(X_{ni}, \theta_A^*) \right| \right) \\ &= V_A^{1/2} \left( \limsup_{n \rightarrow \infty} \left| V_A(P_n)^{-1/2} \frac{1}{n} \sum_{i=1}^n \nabla_{\theta_{A_j}} \ln f_A(X_{ni}, \theta_A^*) \right| \right) \\ &\leq \|V_A\|_F^{1/2} \limsup_{n \rightarrow \infty} \left| V_A(P_n)^{-1/2} \frac{1}{n} \sum_{i=1}^n \nabla_{\theta_{A_j}} \ln f_A(X_{ni}, \theta_A^*) \right| \end{aligned}$$

The summation term in (21) then has two possible behaviors: Either

$$\limsup_{n \rightarrow \infty} \left| \frac{1}{n} \sum_{i=1}^n \nabla_{\theta_{A_j}} \ln f_A(X_{ni}, \theta_A^*) \right| \leq \|V_A\|_F^{1/2} \sqrt{2 \ln n} \quad (22)$$

almost surely for the general triangular array case (by Lemma 7 under Assumption 14 and the fact that  $E_{P_0}[\nabla_{\theta_{A_j}} \ln f_A(X_{ni}, \theta_A^*)] = 0$ ), or

$$\limsup_{n \rightarrow \infty} \left| \frac{1}{n} \sum_{i=1}^n \nabla_{\theta_{A_j}} \ln f_A(X_i, \theta_A^*) \right| \leq \|V_A\|_F^{1/2} \sqrt{2 \ln \ln n} \quad (23)$$

almost surely when  $X_{ni}$  reduces to a sequence ( $X_{ni} = X_i$  and  $V_A(P_n) = V_A$ ), by the Law of Iterated Logarithm (LIL) (Hartman and Wintner (1941)), since Assumption 14 implies existence of the variance. In either case, it follows that<sup>11</sup>

$$\hat{\theta}_A - \theta_A^* = O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right) \quad (24)$$

with  $s = 1$  (for arrays) or  $s = 2$  (for sequences), where  $\ln^{\circ s}$  represents  $s$  application(s) of the  $\ln$  function. A similar result holds for  $\hat{\theta}_B$ .

We now consider each term in the statistic  $\tilde{t}_n = (\hat{\varepsilon}_n \hat{L}_S + \hat{L}_J) / \hat{\sigma}$  where

$$\begin{aligned} \hat{L}_S &:= n^{-1/2} \sum_{i \text{ even}} \ln f_A(X_{ni}, \hat{\theta}_A) - n^{-1/2} \sum_{i \text{ odd}} \ln f_B(X_{ni}, \hat{\theta}_B), \\ \hat{L}_J &:= n^{-1/2} \sum_{i=1}^n \left( \ln f_A(X_{ni}, \hat{\theta}_A) - \ln f_B(X_{ni}, \hat{\theta}_B) \right). \end{aligned}$$

---

<sup>11</sup>For some random sequence  $R_n$  and some deterministic sequence  $r_n$ , we write  $R_n = O_{as}(r_n)$  if and only if there exists a finite  $C$  such that  $P(\limsup_{n \rightarrow \infty} |R_n/r_n| \leq C) = 0$ .

Write  $\hat{\varepsilon}_n \hat{L}_S = \varepsilon_n L_S + (\hat{\varepsilon}_n - \varepsilon_n) L_S + \hat{\varepsilon}_n (R_{\theta_A} - R_{\theta_B})$  with

$$\begin{aligned}\varepsilon_n &:= c_\alpha n^{-1/4} \sqrt{\ln \ln n} \\ L_S &:= n^{-1/2} \sum_{i \text{ even}} \ln f_A(X_{ni}, \theta_A^*) - n^{-1/2} \sum_{i \text{ odd}} \ln f_B(X_{ni}, \theta_B^*) \\ R_{\theta_A} &:= n^{-1/2} \sum_{i \text{ even}} \ln f_A(X_{ni}, \hat{\theta}_A) - n^{-1/2} \sum_{i \text{ even}} \ln f_A(X_{ni}, \theta_A^*) \\ R_{\theta_B} &:= n^{-1/2} \sum_{i \text{ odd}} \ln f_B(X_{ni}, \hat{\theta}_B) - n^{-1/2} \sum_{i \text{ odd}} \ln f_B(X_{ni}, \theta_B^*)\end{aligned}$$

We can bound  $R_{\theta_A}$  (and similarly  $R_{\theta_B}$ ) using an expansion to second order about  $\theta_A = \theta_A^*$ :

$$\begin{aligned}R_{\theta_A} &= n^{-1/2} \sum_{i \text{ even}} \ln f_A(X_{ni}, \hat{\theta}_A) - n^{-1/2} \sum_{i \text{ even}} \ln f_A(X_{ni}, \theta_A^*) \\ &= n^{-1/2} \sum_{i \text{ even}} \ln f_A(X_{ni}, \theta_A^*) + (\hat{\theta}_A - \theta_A^*)' n^{-1/2} \sum_{i \text{ even}} \nabla_{\theta_A} \ln f_A(X_{ni}, \theta_A^*) \\ &\quad + \frac{1}{2} (\hat{\theta}_A - \theta_A^*)' \left( n^{-1/2} \sum_{i \text{ even}} \nabla_{\theta_A}^2 \ln f_A(X_{ni}, \bar{\theta}_A) \right) (\hat{\theta}_A - \theta_A^*) \\ &\quad - n^{-1/2} \sum_{i \text{ even}} \ln f_A(X_{ni}, \theta_A^*) \\ &= (\hat{\theta}_A - \theta_A^*)' n^{1/2} n^{-1} \sum_{i \text{ even}} \nabla_{\theta_A} \ln f_A(X_{ni}, \theta_A^*) \\ &\quad + \frac{n^{1/2}}{4} (\hat{\theta}_A - \theta_A^*)' \left( (n/2)^{-1} \sum_{i \text{ even}} \nabla_{\theta_A}^2 \ln f_A(X_{ni}, \bar{\theta}_A) \right) (\hat{\theta}_A - \theta_A^*)\end{aligned}$$

where  $\bar{\theta}_A$  is a mean value on the line segment joining  $\hat{\theta}_A$  and  $\theta_A^*$ . Then, we use (24) and Lemma 5 applied to  $n^{-1} \sum_{i \text{ even}} \nabla_{\theta_A}^2 \ln f_A(X_{ni}, \bar{\theta}_A)$  under Assumptions 9 and 11:

$$\begin{aligned}\|R_{\theta_A}\| &= O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right) n^{1/2} O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right) \\ &\quad + n^{1/2} O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right) (O(1) + o_{as}(1)) O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right) \\ &= O_{as} \left( n^{-1/2} \ln^{\circ s} n \right)\end{aligned}$$

Next,  $\hat{L}_J = L_J + L_{J2A} - L_{J2B}$  where

$$\begin{aligned}L_J &:= n^{-1/2} \sum_{i=1}^n (\ln f_A(X_{ni}, \theta_A^*) - \ln f_B(X_{ni}, \theta_B^*)) \\ L_{J2A} &:= n^{-1/2} \sum_{i=1}^n \left( \ln f_A(X_{ni}, \hat{\theta}_A) - \ln f_A(X_{ni}, \theta_A^*) \right) \\ L_{J2B} &:= n^{-1/2} \sum_{i=1}^n \left( \ln f_A(X_{ni}, \hat{\theta}_B) - \ln f_B(X_{ni}, \theta_B^*) \right).\end{aligned}$$

The terms  $L_{J2A}$  and  $L_{J2B}$  can be bounded using the same techniques as for  $R_{\theta_A}$  and we have:

$$|L_{J2A}| = O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right)$$

and similarly for  $L_{J2B}$ . Next, let  $\sigma_S^2 := \frac{1}{2}(\sigma_A^2 + \sigma_B^2)$ ,  $\sigma_k^2 := \sigma_k^2(P_0)$ , and  $\hat{\sigma}_S^2 := \frac{1}{2}(\hat{\sigma}_A^2 + \hat{\sigma}_B^2)$ . We have

$$\begin{aligned} \hat{\sigma}_A^2 &= \frac{1}{n} \sum_{i=1}^n \left( \ln f_A \left( X_{ni}, \hat{\theta}_A \right) \right)^2 - \left( \frac{1}{n} \sum_{i=1}^n \ln f_A \left( X_{ni}, \hat{\theta}_A \right) \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left( \ln f_A \left( X_{ni}, \theta_A^* \right) \right)^2 + \left( \hat{\theta}_A - \theta_A^* \right)' \frac{1}{n} \sum_{i=1}^n \ln f_A \left( X_{ni}, \bar{\theta}_A \right) \nabla_{\theta_A} \ln f_A \left( X_{ni}, \bar{\theta}_A \right) \\ &\quad - \left( \frac{1}{n} \sum_{i=1}^n \ln f_A \left( X_{ni}, \theta_A^* \right) + O_{as} \left( n^{-1} \ln^{\circ s} n \right) \right)^2 \\ &= E_{P_0} \left[ \left( \ln f_A \left( X_{ni}, \theta_A^* \right) \right)^2 \right] - E_{P_0} \left( \left[ \ln f_A \left( X_{ni}, \theta_A^* \right) \right] \right)^2 + O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right) \\ &\quad + O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right) \left( O(1) + o_{as}(1) \right) \\ &= \sigma_A^2 + O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right) \end{aligned}$$

where the rate of convergence of the first term follows from Lemma 7 (under Assumption 15) while the one of the second term follows from (24) and Lemma 5 under Assumptions 9 and 12. Similarly, we have  $\hat{\sigma}_B^2 = \sigma_B^2 + O_{as}(n^{-1/2} \sqrt{\ln^{\circ s} n})$  and, thus,  $\hat{\sigma}_S^2 = \sigma_S^2 + O_{as}(n^{-1/2} \sqrt{\ln^{\circ s} n})$ . By a similar reasoning, by Assumptions 16–19, we have  $\hat{H}_k = H_k + O_{as}(n^{-1/2} \sqrt{\ln^{\circ s} n})$  and  $\hat{V}_k = V_k + O_{as}(n^{-1/2} \sqrt{\ln^{\circ s} n})$  for  $k = A, B$ . Below, we will use  $\ln \ln n = O(\ln n)$  to simplify some expressions. From the convergence of  $\hat{\sigma}_S^2$ ,  $\hat{H}_k$  and  $\hat{V}_k$ , it also follows that

$$\begin{aligned} \hat{\varepsilon}_n &= \varepsilon_n + n^{-1/4} \sqrt{\ln \ln n} O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right) = \varepsilon_n + O_{as} \left( n^{-3/4} \sqrt{\ln \ln n} \sqrt{\ln^{\circ s} n} \right) \\ &= \varepsilon_n + O_{as} \left( n^{-3/4} \ln^{\circ s} n \right) \end{aligned}$$

and similarly,

$$\begin{aligned} \hat{\varepsilon}_n^2 &= \left( \varepsilon_n + O_{as} \left( n^{-3/4} \ln^{\circ s} n \right) \right)^2 = \varepsilon_n^2 + O \left( n^{-1/4} \sqrt{\ln \ln n} \right) O_{as} \left( n^{-3/4} \ln^{\circ s} n \right) \\ &= \varepsilon_n^2 + O_{as} \left( n^{-1} (\ln^{\circ s} n)^{3/2} \right). \end{aligned}$$

Next, one can handle  $\hat{\sigma}^2$  by a similar reasoning, invoking Assumptions 9 and 12 and Lemma 5 to yield:

$$\hat{\sigma}^2 = \sigma^2 + O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right).$$

Letting  $\tilde{\sigma}^2(\varepsilon_n) := \varepsilon_n^2 \sigma_S^2 + (1 + \varepsilon_n) \sigma^2$ , we can also write

$$\begin{aligned} \hat{\sigma}^2 &= \varepsilon_n^2 \sigma_S^2 + (1 + \varepsilon_n) \sigma^2 + \varepsilon_n^2 (\hat{\sigma}_S^2 - \sigma_S^2) + (\hat{\varepsilon}_n^2 - \varepsilon_n^2) \hat{\sigma}_S^2 + (\hat{\varepsilon}_n - \varepsilon_n) \sigma^2 + (1 + \hat{\varepsilon}_n) (\hat{\sigma}^2 - \sigma^2) \\ &= \tilde{\sigma}^2(\varepsilon_n) + O \left( n^{-1/2} \ln \ln n \right) O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right) + O_{as} \left( n^{-1} (\ln^{\circ s} n)^{3/2} \right) O_{as}(1) \end{aligned}$$

$$\begin{aligned}
& + O_{as} \left( n^{-3/4} \ln^{\circ s} n \right) O(1) + \left( 1 + O \left( n^{-1/4} \sqrt{\ln \ln n} \right) \right) O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right) \\
& = \tilde{\sigma}^2(\varepsilon_n) + O \left( n^{-1} (\ln \ln n)^{3/2} \right) + O_{as} \left( n^{-1} (\ln^{\circ s} n)^{3/2} \right) + O_{as} \left( n^{-3/4} \ln^{\circ s} n \right) \\
& \quad + O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right) \\
& = \tilde{\sigma}^2(\varepsilon_n) + O_{as} \left( n^{-1/2} \sqrt{\ln^{\circ s} n} \right)
\end{aligned}$$

Collecting all remainder terms for the triangular array case ( $s = 1$ ), we have

$$\begin{aligned}
\tilde{t}_n &= \frac{\hat{\varepsilon}_n \hat{L}_S + \hat{L}_J}{\hat{\sigma}} = \frac{\varepsilon_n L_S + (\hat{\varepsilon}_n - \varepsilon_n) L_S + \hat{\varepsilon}_n (R_{\theta_A} - R_{\theta_B}) + L_J + L_{J2A} - L_{J2B}}{\hat{\sigma}} \\
&= \frac{\varepsilon_n L_S + O_{as} \left( n^{-3/4} \ln n \right) O_{as}(1) + O_{as} \left( n^{-1/4} \sqrt{\ln n} \right) O_{as} \left( n^{-1/2} \ln n \right)}{\tilde{\sigma}(\varepsilon_n) + O_{as} \left( n^{-1/2} \sqrt{\ln n} \right)} \\
&\quad + \frac{L_J + O_{as} \left( n^{-1/2} \ln n \right)}{\tilde{\sigma}(\varepsilon_n) + O_{as} \left( n^{-1/2} \sqrt{\ln n} \right)} \\
&= \frac{\varepsilon_n L_S + L_J + O_{as} \left( n^{-1/2} \ln n \right)}{\tilde{\sigma}(\varepsilon_n) + O_{as} \left( n^{-1/2} \sqrt{\ln n} \right)} = \frac{\varepsilon_n L_S + L_J}{\tilde{\sigma}(\varepsilon_n)} + O_{as} \left( n^{-1/2} \ln n \right),
\end{aligned}$$

that is,  $\tilde{t}_n = t_n + \Delta t_n$  with

$$\begin{aligned}
t_n &:= \frac{\varepsilon_n L_S + L_J}{\tilde{\sigma}(\varepsilon_n)} \\
|\Delta t_n| &\leq \Delta \bar{t}_n \text{ a.s.}
\end{aligned}$$

for  $\Delta \bar{t}_n := B n^{-1/2} \ln n$  for some constant  $B$  and where ‘‘a.s.’’ denotes ‘‘almost surely as  $n \rightarrow \infty$ ’’, i.e., the event  $|\Delta t_n| > \Delta \bar{t}_n$  has probability zero for all  $n \geq n_0$  with  $n_0$  sufficiently large.

When the models are not overlapping, both  $L_S$  and  $L_J$  are asymptotically normal, since they are iid sample averages (evaluated at the true parameter values) of bounded variance quantities. Moreover, by the Berry-Esseen bound (since Assumption 10 implies that the third moments of the log-likelihood function exist and are uniformly bounded), we have that the deviations from normality of finite sample distribution of the normalized statistic  $(\varepsilon_n L_S + L_J)/\tilde{\sigma}(\varepsilon_n)$  are uniformly bounded by  $C n^{-1/2}$  for some universal constant  $C$  (this remains true for triangular arrays, since the constant is independent of the distribution among distributions sharing the same upper bound on the third moments). Let  $\Phi$  and  $\phi$  respectively denote the cdf and pdf of a standard normal. We then have for  $z > 0$  and  $n \geq n_0$ ,

$$\begin{aligned}
|P_n(\tilde{t}_n \leq z) - \Phi(z)| &= |P_n(t_n + \Delta t_n \leq z) - \Phi(z)| \\
&= |P_n(t_n + \Delta t_n \leq z \mid |\Delta t_n| \leq \Delta \bar{t}_n) P_n(|\Delta t_n| \leq \Delta \bar{t}_n) + \\
&\quad + P_n(t_n + \Delta t_n \leq z \mid |\Delta t_n| > \Delta \bar{t}_n) P_n(|\Delta t_n| > \Delta \bar{t}_n) - \Phi(z)| \\
&= |P_n(t_n + \Delta t_n \leq z \mid |\Delta t_n| \leq \Delta \bar{t}_n) \cdot 1 \\
&\quad + P_n(t_n + \Delta t_n \leq z \mid |\Delta t_n| > \Delta \bar{t}_n) \cdot 0 - \Phi(z)|
\end{aligned}$$

$$\begin{aligned}
&= |P_n(t_n + \Delta t_n \leq z \mid |\Delta t_n| \leq \Delta \bar{t}_n) - \Phi(z)| \\
&\leq \sup_{|u| \leq \Delta \bar{t}_n} |P_n(t_n + u \leq z) - \Phi(z)| = \sup_{|u| \leq \Delta \bar{t}_n} |P_n(t_n \leq z - u) - \Phi(z)| \\
&\leq \sup_{|u| \leq \Delta \bar{t}_n} |\Phi(z - u) - \Phi(z)| + Cn^{-1/2} = \sup_{|u| \leq \Delta \bar{t}_n} \phi(z + \bar{u})|u| + Cn^{-1/2} \\
&\leq \sup_{|\bar{u}| \leq \Delta \bar{t}_n} \phi(z + \bar{u}) \Delta \bar{t}_n + Cn^{-1/2} \\
&= \phi(z + o(1)) \Delta \bar{t}_n + Cn^{-1/2} = O(\Delta \bar{t}_n)
\end{aligned} \tag{25}$$

where  $\bar{u}$  is a mean value satisfying  $|\bar{u}| \leq |u| \leq \Delta \bar{t}_n = o(1)$  and by continuity of  $\phi(\cdot)$ , we have  $\phi(z + o(1)) = \phi(z) + o(1)$ . The above display implies that the normal approximation can be used to calculate the power loss, as long as the power loss is not smaller than  $O(\Delta \bar{t}_n) = O(n^{-1/2} \ln n)$ .

We now calculate the power in the nonoverlapping case. Consider a critical value  $z_{1-\alpha} > 0$  and any alternative hypothesis  $\delta \in \mathbb{R}$ . The worst-case power loss due to taking  $\varepsilon_n$  instead of 0 when calculating  $\tilde{t}_n$  is given by

$$\begin{aligned}
\sup_{\delta \in \mathbb{R}} \left| \Phi\left(\frac{\delta}{\tilde{\sigma}(\varepsilon_n)} - z_{1-\alpha}\right) - \Phi\left(\frac{\delta}{\sigma} - z_{1-\alpha}\right) \right| &= \sup_{\delta \in \mathbb{R}} \left| \phi\left(\frac{\delta}{\sigma} - z_{1-\alpha}\right) \frac{\delta}{\sigma^2} \frac{\partial \tilde{\sigma}(\varepsilon)}{\partial \varepsilon} \Bigg|_{\varepsilon=0} \varepsilon_n + O(\varepsilon_n^2) \right| \\
&= \sup_{\delta \in \mathbb{R}} \left| \phi\left(\frac{\delta}{\sigma} - z_{1-\alpha}\right) \frac{\delta \varepsilon_n}{2\sigma} + O(\varepsilon_n^2) \right| \\
&= \frac{1}{2} \phi(z_b) (z_b + z_{1-\alpha}) \varepsilon_n + O(\varepsilon_n^2)
\end{aligned}$$

where  $z_b = \frac{\delta^*}{\sigma} - z_{1-\alpha}$  and  $\delta^* := \frac{\sigma}{2}(z_{1-\alpha} + (4 + z_{1-\alpha}^2)^{1/2})$ . The first and second equality use  $\tilde{\sigma}(\varepsilon) = \sigma$  for  $\varepsilon = 0$  and

$$\frac{\partial \tilde{\sigma}(\varepsilon)}{\partial \varepsilon} \Bigg|_{\varepsilon=0} = \frac{\partial}{\partial \varepsilon} \sqrt{\varepsilon^2 \sigma_S^2 + (1 + \varepsilon) \sigma^2} \Bigg|_{\varepsilon=0} = \frac{1}{2} \frac{(2\varepsilon \sigma_S^2 + \sigma^2)}{\sqrt{\varepsilon^2 \sigma_S^2 + (1 + \varepsilon) \sigma^2}} \Bigg|_{\varepsilon=0} = \frac{\sigma}{2}.$$

The third equality above follows by taking the derivative of  $\phi\left(\frac{\delta}{\sigma} - z_{1-\alpha}\right) \frac{\delta \varepsilon_n}{2\sigma}$  with respect to  $\delta$ , setting it to zero and noticing that the solution  $\delta^*$  is, in fact, the global maximum. Substituting in  $\varepsilon_n$  then gives

$$\sup_{\delta \in \mathbb{R}} \left| \Phi\left(\frac{\delta}{\tilde{\sigma}(\varepsilon_n)} - z_{1-\alpha}\right) - \Phi\left(\frac{\delta}{\sigma} - z_{1-\alpha}\right) \right| = Mn^{-1/4} \sqrt{\ln \ln n} + o\left(n^{-1/4} \sqrt{\ln \ln n}\right).$$

We now calculate the size distortion when the models are overlapping. In the overlapping case, we need to provide a more precise bound on the remainder terms of  $\hat{L}_J = L_J + L_{J2A} - L_{J2B}$ , because the leading term vanishes ( $L_J = 0$ ) due to the overlap. Drifting sequences of models are not needed for the size calculation, so the triangular array  $X_{ni}$  can be replaced by a simple iid sequence  $X_i$  drawn from  $P_0$ . Letting  $\hat{g}_A := \frac{1}{n} \sum_{i=1}^n (\nabla_{\theta_A} \ln f_A(X_i, \theta_A^*))$ , we have

$$\begin{aligned}
L_{J2A} &= \frac{n^{1/2}}{2} \hat{g}'_A (H_A^{-1} + o_{as}(1)) \hat{g}_A \\
&= \frac{n^{1/2}}{2} \hat{g}'_A V_A^{-1/2} V_A^{1/2} (H_A^{-1} + o_{as}(1)) V_A^{1/2} V_A^{-1/2} \hat{g}_A \\
&= -\frac{n^{1/2}}{2} Z'_A V_A^{1/2} (-H_A^{-1} + o_{as}(1)) V_A^{1/2} Z_A
\end{aligned}$$

where  $Z_A := V_A^{-1/2} \hat{g}_A$ . The matrix  $V_A^{1/2} (-H_A)^{-1} V_A^{1/2}$  is symmetric so it is diagonalizable, with eigenvalues  $\lambda_j$  and orthogonal eigenvectors  $v_j$  (normalized to  $\|v_j\| = 1$ ). Moreover, the eigenvalues are all positive (because both  $-H_A$  and  $V_A$  are positive-definite) and we can write  $V^{1/2} (-H)^{-1} V^{1/2} = \sum_{j=1}^{\dim \theta_A} v_j \lambda_j v_j'$  and thus:

$$\begin{aligned} |L_{J2A}| &= -L_{J2A} = \frac{n^{1/2}}{2} Z_A' \left( \sum_{j=1}^{\dim \theta_A} v_j \lambda_j v_j' + o_{as}(1) \right) Z_A. \\ &= \frac{n^{1/2}}{2} \sum_{j=1}^{\dim \theta_A} Z_A' v_j \lambda_j v_j' Z_A + o_{as}(1) \frac{n^{1/2}}{2} Z_A' Z_A \\ &= \frac{n^{1/2}}{2} \sum_{j=1}^{\dim \theta_A} \lambda_j (v_j' Z_A)^2 + o_{as}(1) \frac{n^{1/2}}{2} Z_A' Z_A \end{aligned}$$

By construction, the covariance matrix of the  $v_j' Z_A$  is the identity matrix  $I$ . We can then use the Law of the Iterated Logarithm (Hartman and Wintner (1941)) to conclude  $|v_j' Z_A| \leq n^{-1/2} \sqrt{2 \ln \ln n}$  almost surely. We then have

$$\begin{aligned} |L_{J2A}| &\leq \frac{n^{1/2}}{2} \sum_{j=1}^{\dim \theta_A} \lambda_j \left( n^{-1/2} \sqrt{2 \ln \ln n} \right)^2 + o_{as}(1) \frac{n^{1/2}}{2} (\dim \theta_A) \left( n^{-1/2} \sqrt{2 \ln \ln n} \right)^2 \\ &= \left( n^{-1/2} \ln \ln n \right) \sum_{j=1}^{\dim \theta_A} \lambda_j + o_{as} \left( n^{-1/2} \ln \ln n \right) \\ &= \left( n^{-1/2} \ln \ln n \right) \text{tr} \left( V_A^{1/2} (-H_A)^{-1} V_A^{1/2} \right) + o_{as} \left( n^{-1/2} \ln \ln n \right) \\ &= \left| \text{tr} (H_A^{-1} V_A) \right| \left( n^{-1/2} \ln \ln n \right) + o_{as} \left( n^{-1/2} \ln \ln n \right) \end{aligned}$$

A similar reasoning holds for  $|L_{J2B}|$  and since both  $L_{J2A}$  and  $L_{J2B}$  have the same sign and  $L_J = 0$ , we have

$$\begin{aligned} \left| \hat{L}_J \right| &= |L_J + L_{J2A} - L_{J2B}| = |L_{J2A} - L_{J2B}| \\ &\leq \max \{ |L_{J2A}|, |L_{J2B}| \} \leq \max \{ \left| \text{tr} (H_A^{-1} V_A) \right|, \left| \text{tr} (H_B^{-1} V_B) \right| \} n^{-1/2} \ln \ln n \text{ a.s.} \\ &= \Lambda n^{-1/2} \ln \ln n, \end{aligned}$$

where  $\Lambda := \max \{ \left| \text{tr} (H_A^{-1} V_A) \right|, \left| \text{tr} (H_B^{-1} V_B) \right| \}$ . In the overlapping case,  $\tilde{\sigma}^2(\varepsilon) = \varepsilon^2 \sigma_S^2 + (1 + \varepsilon) \sigma_J^2 = \varepsilon^2 \sigma_S^2$  since  $\sigma_J^2 = 0$ . We can now compute the worst-case size distortion in  $\tilde{t}_n$ . Collecting the order of all remainders, we have,

$$\begin{aligned} \tilde{t}_n &= \frac{\hat{\varepsilon}_n \hat{L}_S + \hat{L}_J}{\hat{\sigma}} = \frac{\varepsilon_n L_S + (\hat{\varepsilon}_n - \varepsilon_n) L_S + \hat{\varepsilon}_n (R_{\theta_A} - R_{\theta_B}) + L_J + L_{J2A} - L_{J2B}}{\tilde{\sigma}(\varepsilon_n) + O_{as} \left( n^{-1/2} \sqrt{\ln \ln n} \right)} \\ &= \frac{\varepsilon_n L_S + O_{as} \left( n^{-3/4} \ln \ln n \right) O_{as}(1) + O_{as} \left( n^{-1/4} \sqrt{\ln \ln n} \right) O_{as} \left( n^{-1/2} \ln \ln n \right)}{\varepsilon_n \sigma_S + O_{as} \left( n^{-1/2} \sqrt{\ln \ln n} \right)} \end{aligned}$$

$$\begin{aligned}
& + \frac{L_{J2A} - L_{J2B}}{\varepsilon_n \sigma_S + O_{as} \left( n^{-1/2} \sqrt{\ln \ln n} \right)} \\
& = \frac{\varepsilon_n L_S + (L_{J2A} - L_{J2B}) + O_{as} \left( n^{-3/4} (\ln \ln n)^{3/2} \right)}{\varepsilon_n \sigma_S + O_{as} \left( n^{-1/2} \sqrt{\ln \ln n} \right)} \\
& = \frac{L_S}{\sigma_S} \frac{\varepsilon_n + (L_{J2A} - L_{J2B}) / L_S + O_{as} \left( n^{-3/4} (\ln \ln n)^{3/2} \right)}{\varepsilon_n + O_{as} \left( n^{-1/2} \sqrt{\ln \ln n} \right)} \\
& = \frac{L_S}{\sigma_S} \frac{1 + (L_{J2A} - L_{J2B}) / (\varepsilon_n L_S) + O_{as} \left( n^{-3/4} (\ln \ln n)^{3/2} / \varepsilon_n \right)}{1 + O_{as} \left( n^{-1/2} \left( \sqrt{\ln \ln n} \right) / \varepsilon_n \right)} \\
& = \left( \frac{L_S}{\sigma_S} + \frac{(L_{J2A} - L_{J2B})}{\varepsilon_n \sigma_S} + O_{as} \left( n^{-3/4} (\ln \ln n)^{3/2} / \varepsilon_n \right) \right) \times \\
& \quad \times \frac{1}{\left( 1 + O_{as} \left( n^{-1/2} \left( \sqrt{\ln \ln n} \right) / \varepsilon_n \right) \right)} \\
& = \frac{L_S}{\sigma_S} + \frac{(L_{J2A} - L_{J2B})}{\varepsilon_n \sigma_S} + O_{as} \left( n^{-1/2} \sqrt{\ln \ln n} / \varepsilon_n \right) \\
& = \frac{L_S}{\sigma_S} + \frac{(L_{J2A} - L_{J2B})}{\varepsilon_n \sigma_S} + O_{as} \left( n^{-1/2} \sqrt{\ln \ln n} / \left( n^{-1/4} \sqrt{\ln \ln n} \right) \right) \\
& = \frac{L_S}{\sigma_S} + \Delta t_n
\end{aligned}$$

where  $\Delta t_n := (L_{J2A} - L_{J2B}) / (\varepsilon_n \sigma_S) + O_{as}(n^{-1/4})$ . We can bound  $\Delta t_n$  as follows, substituting in  $\varepsilon_n$ :

$$\begin{aligned}
|\Delta t_n| & = \frac{|L_{J2A} - L_{J2B}|}{\varepsilon_n \sigma_S} + O_{as} \left( n^{-1/4} \right) \\
& \leq \frac{\Lambda n^{-1/2} \ln \ln n}{\varepsilon_n \sigma_S} + O_{as} \left( n^{-1/4} \right) \quad \text{a.s.} \\
& = \frac{\Lambda n^{-1/2} \ln \ln n}{\sigma_S c_\alpha n^{-1/4} \sqrt{\ln \ln n}} + O_{as} \left( n^{-1/4} \right) \\
& = B n^{-1/4} \sqrt{\ln \ln n} + O_{as} \left( n^{-1/4} \right)
\end{aligned}$$

where  $B := \Lambda / (\sigma_S c_\alpha)$ . The size distortion for a given critical value  $z_{1-\alpha}$  is then given by a similar calculation as in (25) with  $\Delta \tilde{t}_n = B n^{-1/4} \sqrt{\ln \ln n} + \tilde{B} n^{-1/4}$  with sufficiently large  $\tilde{B}$ :

$$|P_0(\tilde{t}_n \leq z_{1-\alpha}) - \Phi(z_{1-\alpha})| = \phi(z_{1-\alpha}) B n^{-1/4} \sqrt{\ln \ln n} + C n^{-1/2} + O(n^{-1/4}).$$

for some universal constant  $C$ . Noticing that  $M = \phi(z_{1-\alpha}) B$  then yields the desired result.

Finally, we observe that  $\varepsilon_n$  is in  $\mathcal{E}$  by construction and since we have shown that  $\hat{\varepsilon}_n = \varepsilon_n + O_{as}(n^{-3/4} \ln^{os} n)$ , we automatically have  $\hat{\varepsilon}_n - \varepsilon_n = O_p(n^{-1/2})$ , for either sequences ( $s = 2$ ) or triangular arrays ( $s = 1$ ), and it follows that  $\hat{\varepsilon}_n$  satisfies Assumptions 7 and 8. Q.E.D.



## C Auxiliary Lemmas

The following Lemma provides a uniform strong law of large numbers for triangular arrays. It is stated for scalars, but can also be used, element by element, for vectors valued  $g(x, \theta)$ .

**Lemma 5.** *For  $n \in \mathbb{N}$ , let  $X_{ni}$  for  $i = 1, \dots, n$  be iid random variables taking value in  $\mathbb{R}^{d_x}$  and drawn from the probability measure  $P_n$ . Assume that the measures  $P_n$  converge weakly to some measure  $P_0$  and that each  $P_n(x)$  admits a Radon-Nikodym derivative  $p_n(x)$  with respect to  $P_0(x)$ . For  $\Theta$  compact (under some metric  $d_\theta(\cdot, \cdot)$ ), let  $g : \mathbb{R}^{d_x} \times \Theta \mapsto \mathbb{R}$  be continuous in  $x$  at each  $\theta \in \Theta$ . Assume further that there exists  $G(x)$  such that  $E_{P_0}[G(X_{0i})] < \infty$  (for  $X_{0i}$  drawn from  $P_0$ ) and such that, for all  $\theta \in \Theta$  and  $n \in \mathbb{N}$ ,*

$$|g(x, \theta)| p_n(x) \leq G(x)$$

and that there exists  $\bar{G} < \infty$  such that  $E_{P_n}[|g(X_{ni}, \theta)|^4] \leq \bar{G}$  for all  $i = 1, \dots, n$ , all  $n \in \mathbb{N}$  and all  $\theta \in \Theta$ . Then,

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n g(X_{ni}, \theta) - g(\theta) \right| \xrightarrow{as} 0$$

for  $g(\theta) := E_{P_0}[g(X_{0i}, \theta)]$ , where  $X_{0i}$  is drawn from  $P_0$ .

*Proof.* This proof parallels the one of Lemma 1 in [Tauchen \(1985\)](#), but adapted for triangular arrays. Define

$$u(x, \theta, d) = \sup_{\tilde{\theta}: d_\theta(\tilde{\theta}, \theta) \leq d} \left| g(x, \tilde{\theta}) - g(x, \theta) \right|.$$

By almost sure continuity of  $g(x, \theta)$ ,  $\lim_{d \rightarrow 0} u(x, \theta, d) = 0$  almost surely, for a given  $\theta$ . Also observe that, by  $P_n$  converging weakly to  $P_0$ , we must have that  $p_n(x) \rightarrow 1$  pointwise for all  $x$  in a set of probability 1 under  $P_0$ . To study the convergence of  $E_{P_n}[u(X, \theta, d)]$  as  $d \rightarrow 0$  and  $n \rightarrow \infty$ , we employ dominated convergence. We have

$$E_{P_n}[u(X, \theta, d)] = \int u(x, \theta, d) dP_n(x) = \int u(x, \theta, d) p_n(x) dP_0(x)$$

where

$$|u(x, \theta, d) p_n(x)| \leq \sup_{d_\infty(\tilde{\theta}, \theta) \leq d} \left| g(x, \tilde{\theta}) \right| p_n(x) + |g(x, \theta)| p_n(x) \leq G(x) + G(x) = 2G(x),$$

where  $\int G(x) dP_0(x) < \infty$ . Thus, for a given  $\varepsilon > 0$ , there exists  $\bar{d}(\theta)$  and  $\bar{N}(\theta, \varepsilon)$  such that  $E_{P_n}[u(X_{ni}, \theta, d)] \leq \varepsilon$  whenever  $d \leq \bar{d}(\theta)$  and  $n \geq \bar{N}(\theta, \varepsilon)$ . By a similar reasoning,  $|g(\tilde{\theta}) - g(\theta)| \leq \varepsilon$  whenever  $d(\tilde{\theta}, \theta) \leq \bar{d}(\theta)$ . Let  $B(\theta)$  be the open ball of radius  $\bar{d}(\theta)$  about  $\theta$ . By compactness of  $\Theta$ , there exists a finite covering

$B_k = B(\theta_k)$ ,  $k = 1, \dots, K$ . Let  $d_k = \bar{d}(\theta_k)$  and  $\mu_k = E[u(X, \theta_k, d_k)]$  and write, for  $\theta \in B_k$ ,

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^n g(X_{ni}, \theta) - g(\theta) \right| &\leq \left| \frac{1}{n} \sum_{i=1}^n g(X_{ni}, \theta) - g(X_{ni}, \theta_k) \right| + \left| \frac{1}{n} \sum_{i=1}^n g(X_{ni}, \theta_k) - E_{P_0} [g(X_{0i}, \theta_k)] \right| \\ &\quad + |E_{P_0} [g(X_{0i}, \theta_k)] - g(\theta)| \\ &\leq \left| \frac{1}{n} \sum_{i=1}^n u(X_{ni}, \theta_k, d_k) - \mu_k \right| + \mu_k + \left| \frac{1}{n} \sum_{i=1}^n g(X_{ni}, \theta_k) - E_{P_0} [g(X_{0i}, \theta_k)] \right| \\ &\quad + |g(\theta_k) - g(\theta)| \\ &:= R_1 + \mu_k + R_2 + |g(\theta_k) - g(\theta)| \end{aligned}$$

By construction,  $\mu_k \leq \varepsilon$  and  $|g(\theta_k) - g(\theta)| \leq \varepsilon$  for all  $n \geq \bar{N}(\theta_k, \varepsilon)$ . To apply a strong law of large number for triangular arrays (Lemma 6) for  $R_1$  and  $R_2$  above, we need to calculate fourth moments of the summands. We have

$$\begin{aligned} E \left[ |g(X_{ni}, \theta_k) - E_{P_0} [g(X_{0i}, \theta_k)]|^4 \right] &\leq 8 \left( E \left[ |g(X_{ni}, \theta_k)|^4 \right] + |E_{P_0} [g(X_{0i}, \theta_k)]|^4 \right) \\ &\leq 16E \left[ |g(X_{ni}, \theta)|^4 \right] \leq 16\bar{G} \end{aligned}$$

by the  $C_r$  and Jensen's inequalities and by the uniform boundedness of the fourth moment assumption. Similarly,

$$E \left[ |u(X_{ni}, \theta_k, d_k)|^4 \right] = E \left[ \left| \sup_{\bar{\theta}: d_{\bar{\theta}}(\bar{\theta}, \theta_k) \leq d_k} |g(X_{ni}, \bar{\theta}) - g(X_{ni}, \theta)| \right|^4 \right] = E \left[ |g(X_{ni}, \theta^*) - g(X_{ni}, \theta)|^4 \right]$$

for some  $\theta^*$ , by compactness of (the closure of)  $B(\theta_k)$ . By the  $C_r$  inequality, we have  $E[|g(x, \theta^*) - g(x, \theta)|^4] \leq 16\bar{G}$ . Hence, we can apply Lemma 6 to conclude that there exists  $N_k(\varepsilon)$  such that  $R_1 \leq \varepsilon$  and  $R_2 \leq \varepsilon$  almost surely for all  $n \geq N_k(\varepsilon)$ . Thus,

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n g(X_{ni}, \theta) - g(\theta) \right| \leq 4\varepsilon$$

for  $n \geq \max_k \max\{N_k(\varepsilon), \bar{N}(\theta_k, \varepsilon)\}$  almost surely. Since  $\varepsilon$  was arbitrary, the conclusion follows. Q.E.D.

The following lemma is a strong law of large number for triangular arrays.

**Lemma 6.** *Let  $Y_{ni}$  be a triangular array ( $n \in \mathbb{N}$ ,  $i = 1, \dots, n$ ) of random variables, iid across  $i = 1, \dots, n$ . If, for all  $n \in \mathbb{N}$ ,  $i = 1, \dots, n$ ,  $E[Y_{ni}] = 0$  and  $E[|Y_{ni}|^4] \leq \bar{Y} < \infty$ , then  $n^{-1} \sum_{i=1}^n Y_{ni} \xrightarrow{a.s.} 0$ .*

*Proof.* The principle of this proof is borrowed from Example 5.41 in Romano and Siegel (1986). Note that

$$P \left[ \left| \frac{1}{n} \sum_{i=1}^n Y_{ni} \right| \geq \varepsilon \right] \leq \frac{E \left[ \left| \frac{1}{n} \sum_{i=1}^n Y_{ni} \right|^4 \right]}{\varepsilon^4}$$

where

$$\begin{aligned}
E \left[ \left( \frac{1}{n} \sum_{i=1}^n Y_{ni} \right)^4 \right] &= n^{-4} \sum_{i_1=1}^n \sum_{i_2=1}^n \sum_{i_3=1}^n \sum_{i_4=1}^n E [Y_{ni_1} Y_{ni_2} Y_{ni_3} Y_{ni_4}] \\
&= n^{-4} \sum_{i_1=1}^n \sum_{i_2=1}^n E \left[ |Y_{ni_1}|^2 |Y_{ni_2}|^2 \right] + n^{-4} \sum_{i_1=1}^n E \left[ |Y_{ni_1}|^4 \right] \\
&= n^{-2} E \left[ |Y_{ni}|^2 \right] E \left[ |Y_{ni}|^2 \right] + n^{-3} E \left[ |Y_{ni}|^4 \right] \\
&\leq n^{-2} E \left[ |Y_{ni}|^4 \right]^{1/2} \left( E \left[ |Y_{ni}|^4 \right] \right)^{1/2} + n^{-3} E \left[ |Y_{ni}|^4 \right] \\
&\leq n^{-2} \bar{Y} + n^{-3} \bar{Y}.
\end{aligned}$$

Hence,

$$\sum_{n=1}^{\infty} P \left[ \left| \frac{1}{n} \sum_{i=1}^n Y_{ni} \right| \geq \varepsilon \right] \leq \bar{Y} \sum_{n=1}^{\infty} n^{-2} + \bar{Y} \sum_{n=1}^{\infty} n^{-3} < \infty$$

and, by the Borel-Cantelli Lemma, the event  $\left| n^{-1} \sum_{i=1}^n Y_{ni} \right| \geq \varepsilon$  occurs finitely often almost surely for any  $\varepsilon > 0$ , i.e.  $n^{-1} \sum_{i=1}^n Y_{ni} \xrightarrow{a.s.} 0$ . Q.E.D.

The following provides a law of the “iterated” logarithm for triangular arrays.

**Lemma 7.** *Let  $Y_{ni}$  be a triangular array ( $n \in \mathbb{N}$ ,  $i = 1, \dots, n$ ) of random variables, iid across  $i = 1, \dots, n$ . If, for all  $n \in \mathbb{N}$ ,  $i = 1, \dots, n$ ,  $E[Y_{ni}] = 0$ ,  $E[Y_{ni}^2] > 0$  and  $E[|Y_{ni}|^{4+\delta}] \leq \bar{Y} < \infty$ , then*

$$P \left[ \limsup_{n \rightarrow \infty} \frac{|\sum_{i=1}^n Y_{ni}|}{\sqrt{2E[Y_{ni}^2]n \ln n}} \rightarrow 1 \right] = 1. \tag{26}$$

*Proof.* We use Theorem 1 in [Rubin and Sethuraman \(1965\)](#), in the special case of iid variables across the  $i$  dimension, noting that our assumptions imply their Assumptions (7), (8), (9) and (11) for their  $N$  set to  $n$  and their constants  $q$  and  $c$  set to  $q = 4 + \delta$  and  $c^2 = 2 + \varepsilon$  for any  $\varepsilon < \delta$ . Their Theorem 1 then shows that

$$s_n := P \left[ \left| \sum_{i=1}^n Y_{ni} \right| > c \sqrt{E[Y_{ni}^2]n \ln n} \right] = (1 + o(1)) \frac{n^{-c^2/2}}{c\sqrt{2\pi \ln n}},$$

which can be used with the Borel-Cantelli Lemma. Indeed, the  $s_n$  for  $c^2 = 2 + \varepsilon$  are such that  $\sum_{n=2}^{\infty} s_n < \infty$  for any  $\varepsilon > 0$  since

$$\sum_{n=2}^{\infty} \frac{n^{-1} n^{-\varepsilon/2}}{(\sqrt{2+\varepsilon}) \sqrt{2\pi \ln n}} \leq C \sum_{n=2}^{\infty} n^{-1-\varepsilon/2} < \infty$$

for some universal constant  $C$  and for any  $\varepsilon > 0$ . It follows that the event

$$\left\{ n^{-1} \sum_{i=1}^n Y_{ni} > \sqrt{2+\varepsilon} E[Y_{ni}^2] n^{-1/2} \sqrt{\ln n} \right\}$$

occurs only finitely often for any  $\varepsilon > 0$  arbitrarily close to 0. By a similar reasoning,  $\sum_{n=2}^{\infty} s_n \rightarrow \infty$  for  $\varepsilon < 0$  and that event occurs infinitely often for any  $\varepsilon < 0$  arbitrarily close to 0 and the conclusion (26)

follows. (See also Theorem 3 in [Hu and Weber \(1992\)](#) for a similar use of this inequality, in a context where independence across  $n$  is also assumed, although it is not needed for the application of Theorem 1 in [Rubin and Sethuraman \(1965\)](#).) Q.E.D.

## 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.
- 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.
- HALL, A. R., AND D. PELLETIER (2011): “Nonnested Testing in Models Estimated Via Generalized Method of Moments,” *Econometric Theory*, 27(02), 443–456.
- HANSEN, L. P. (1982): “Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica*, 50(4), 1029–1054.
- HARTMAN, P., AND A. WINTNER (1941): “On the Law of the Iterated Logarithm,” *American Journal of Mathematics*, 63(1), 169–176.
- 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.
- HU, T.-C., AND N. C. WEBER (1992): “On the rate of convergence in the strong law of Large numbers for arrays,” *Bulletin of the Australian Mathematical Society*, 45, 479–482.
- 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.
- LEHMANN, E. L., AND J. P. ROMANO (2005): *Testing Statistical Hypotheses*. Springer, New York.
- MIZON, G. E., AND J.-F. RICHARD (1986): “The Encompassing Principle and its Application to Testing Non-Nested Hypotheses,” *Econometrica*, 54(3), 657–678.
- NEWBY, W. K., AND D. MCFADDEN (1994): “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. IV, pp. 2111–2245. Elsevier Science B.V.

- NEWKEY, W. K., AND R. J. SMITH (2004): “Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators,” *Econometrica*, 72(1), 219–255.
- 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.
- RIVERS, D., AND Q. H. VUONG (2002): “Model Selection Tests for Nonlinear Dynamic Models,” *Econometrics Journal*, 5, 1–39.
- 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.
- ROMANO, J. P., AND A. F. SIEGEL (1986): *Counterexamples in Probability And Statistics*. CRC Press, New York.
- RUBIN, H., AND J. SETHURAMAN (1965): “Probabilities of Moderate Deviations,” *Sankhya A*, 27, 325–346.
- 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. (2013): “A Nondegenerate Vuong Test,” Discussion paper.
- 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.

- TAUCHEN, G. (1985): “Diagnostic Testing And Evaluation Of Maximum Likelihood Models,” *Journal of Econometrics*, 30, 415–443.
- 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.



our test							
n	no reg	$\varepsilon_n = 0.5$	$\varepsilon_n = 1$	optimal	Vuong	Shi	NP
bivariate normal location							
100	0.000	0.041	0.045	0.028	0.000	0.000	
200	0.000	0.046	0.045	0.035	0.000	0.000	
500	0.000	0.039	0.037	0.038	0.000	0.000	
misspecified normals							
100	0.062	0.073	0.076	0.068	0.062	0.051	
200	0.062	0.053	0.059	0.058	0.062	0.044	
500	0.059	0.062	0.062	0.062	0.059	0.044	
correctly specified normals							
100	0.003	0.035	0.039	0.018	0.003	0.000	
200	0.000	0.043	0.045	0.032	0.000	0.000	
500	0.000	0.036	0.034	0.033	0.000	0.000	
nested regressions with one additional regressor							
100	0.001	0.039	0.044	0.040	0.001	0.000	
200	0.000	0.047	0.052	0.047	0.000	0.000	
500	0.000	0.056	0.056	0.056	0.000	0.000	
nested regressions with two additional regressors							
100	0.008	0.049	0.050	0.049	0.006	0.000	0.063
200	0.003	0.049	0.049	0.047	0.002	0.000	0.054
500	0.002	0.059	0.058	0.059	0.002	0.000	0.045

Table 1: Null rejection probabilities of our, Vuong’s, Shi’s, and the Neyman Pearson (‘NP’) test for the different examples and different sample sizes (‘n’). ‘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 (6).

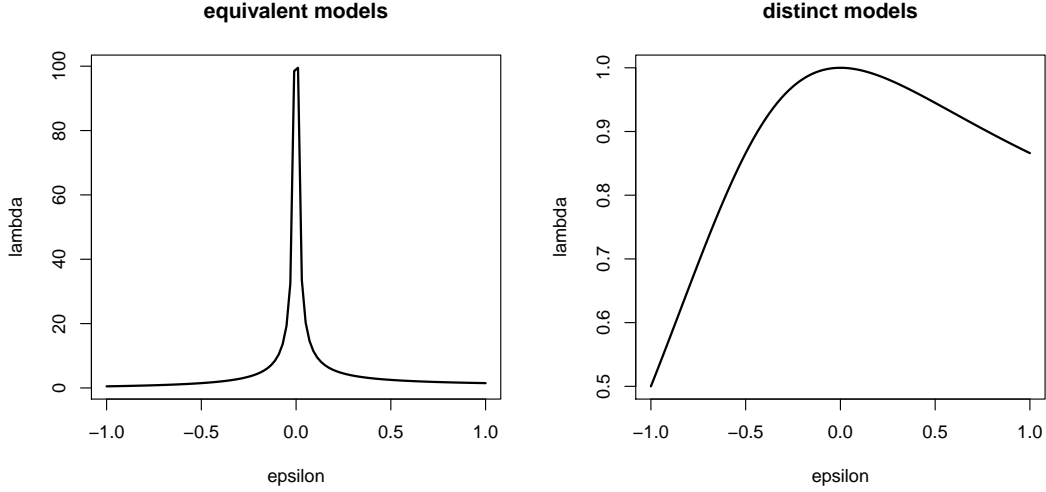


Figure 1: Plots of the noncentrality parameter  $\tilde{\lambda}$  as a function of  $\epsilon$ , when  $\sigma_A^2 = \sigma_B^2 = 1$ ,  $\delta = 1$ , and the models are either equivalent ( $\sigma_{AB} = 1 \Rightarrow \sigma^2 = 0$ ) or distinct ( $\sigma_{AB} = 0.5 \Rightarrow \sigma^2 \neq 0$ ).

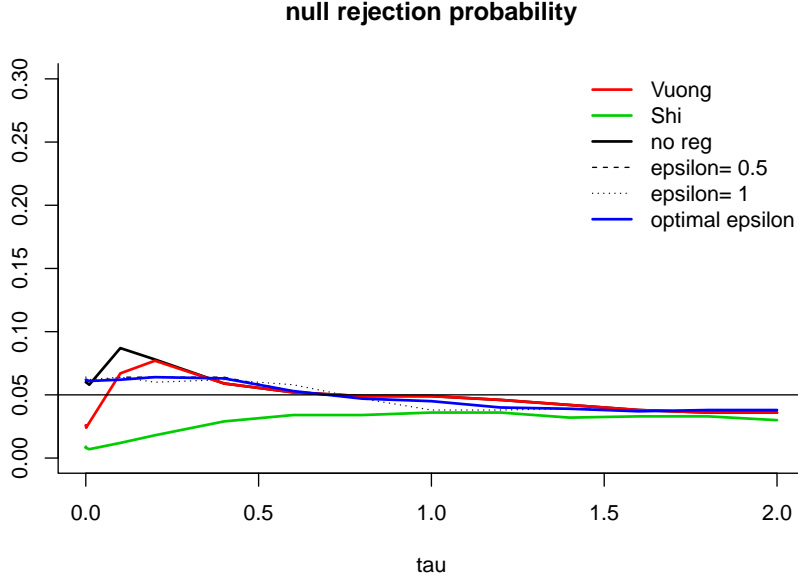


Figure 2: Example 4 (nonnested regressions): Null rejection probabilities of Vuong's, Shi's, and our test. 'no reg', 'epsilon=0.5', 'epsilon=1', and 'optimal epsilon' refer to our test using  $\hat{\epsilon}_n = 0$ ,  $\hat{\epsilon}_n = 0.5$ ,  $\hat{\epsilon}_n = 1$ , and the optimal epsilon in (6).

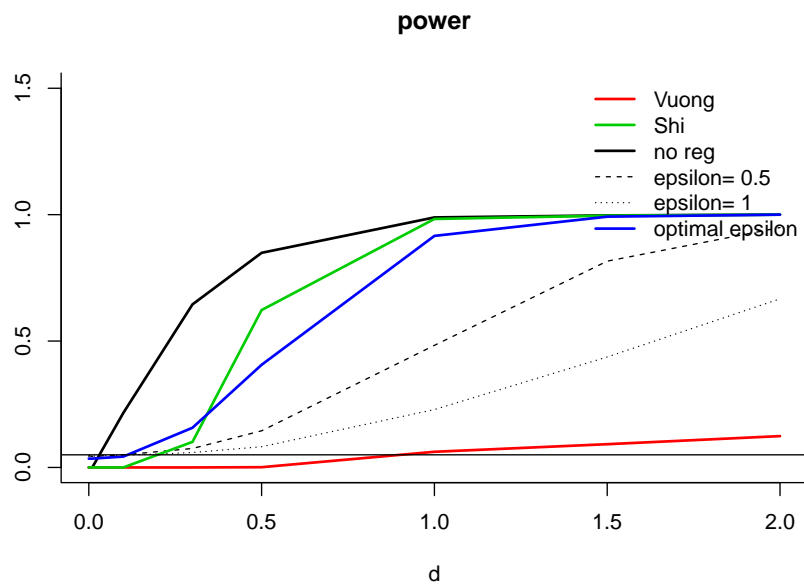


Figure 3: Example 1 (bivariate normal location model,  $\alpha = 0.05$ ): Power curves of Vuong's, Shi's, and our test. 'no reg', 'epsilon=0.5', 'epsilon=1', and 'optimal epsilon' refer to our test using  $\hat{\epsilon}_n = 0$ ,  $\hat{\epsilon}_n = 0.5$ ,  $\hat{\epsilon}_n = 1$ , and the optimal epsilon in (6).

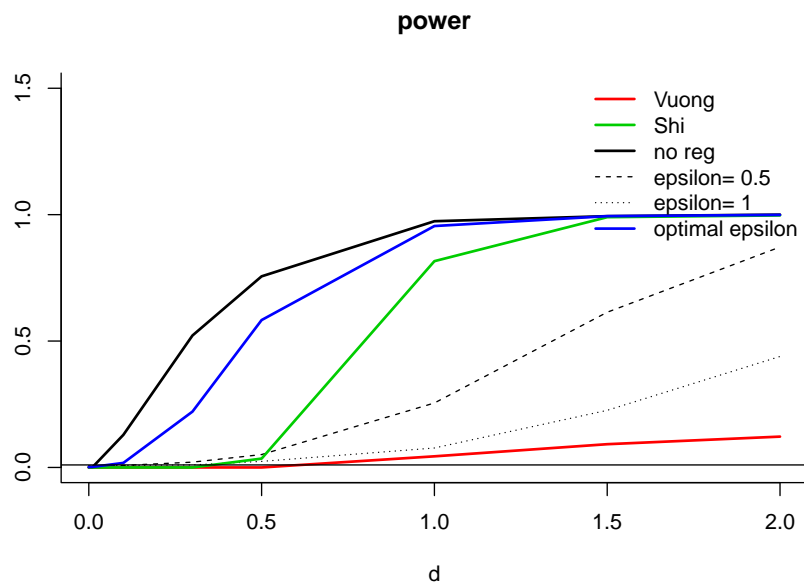


Figure 4: Example 1 (bivariate normal location model,  $\alpha = 0.01$ ): Power curves of Vuong's, Shi's, and our test. 'no reg', 'epsilon=0.5', 'epsilon=1', and 'optimal epsilon' refer to our test using  $\hat{\epsilon}_n = 0$ ,  $\hat{\epsilon}_n = 0.5$ ,  $\hat{\epsilon}_n = 1$ , and the optimal epsilon in (6).

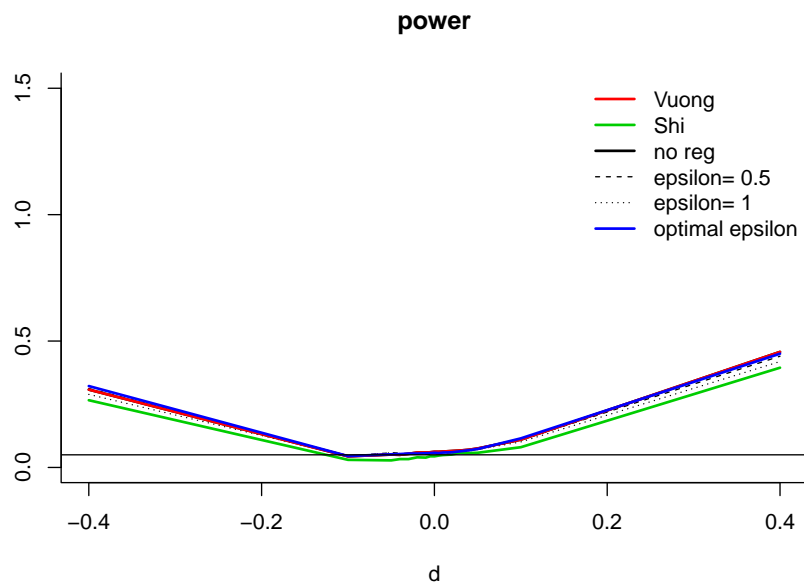


Figure 5: Example 2 (misspecified normals): Power curves of Vuong's, Shi's, and our test. 'no reg', 'epsilon=0.5', 'epsilon=1', and 'optimal epsilon' refer to our test using  $\hat{\varepsilon}_n = 0$ ,  $\hat{\varepsilon}_n = 0.5$ ,  $\hat{\varepsilon}_n = 1$ , and the optimal epsilon in (6).

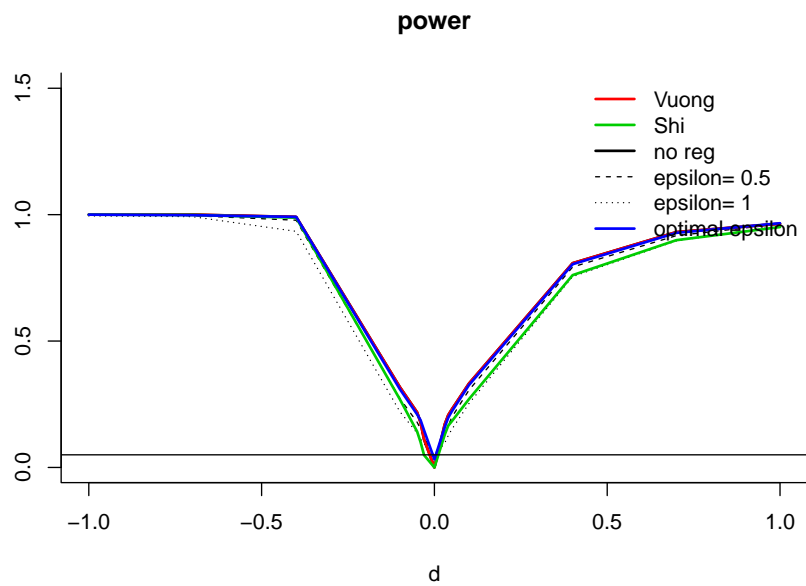


Figure 6: Example 3 (correctly specified normals): Power curves of Vuong's, Shi's, and our test. 'no reg', 'epsilon=0.5', 'epsilon=1', and 'optimal epsilon' refer to our test using  $\hat{\varepsilon}_n = 0$ ,  $\hat{\varepsilon}_n = 0.5$ ,  $\hat{\varepsilon}_n = 1$ , and the optimal epsilon in (6).

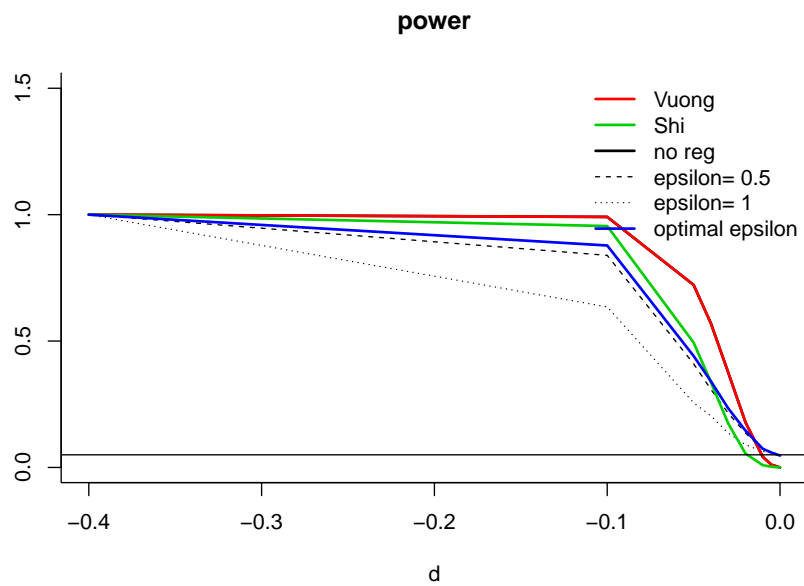


Figure 7: Example 5 (nested regressions with one additional regressor): Power curves of Vuong's, Shi's, and our test. 'no reg', 'epsilon=0.5', 'epsilon=1', and 'optimal epsilon' refer to our test using  $\hat{\epsilon}_n = 0$ ,  $\hat{\epsilon}_n = 0.5$ ,  $\hat{\epsilon}_n = 1$ , and the optimal epsilon in (6).

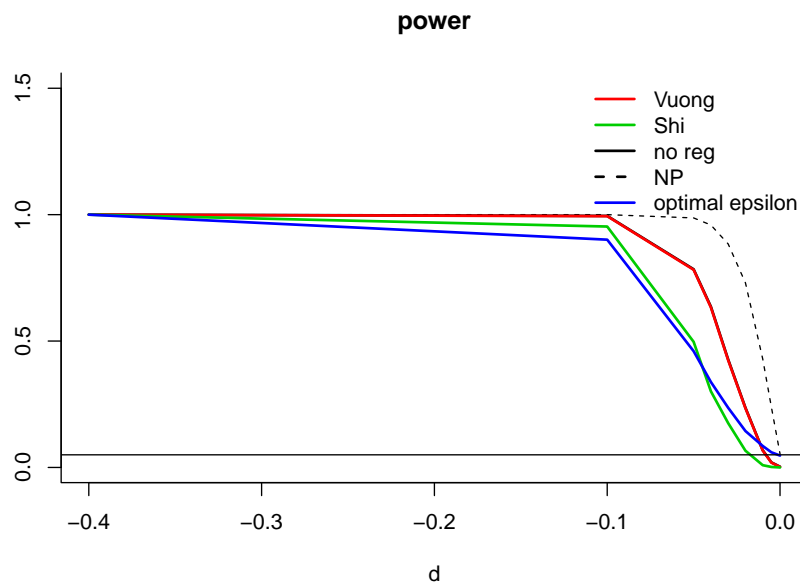


Figure 8: Example 6 (nested regressions with two additional regressors): Power curves of Vuong's, Shi's, and our test. 'NP' refers to the Neyman-Pearson likelihood ratio test, and 'no reg' and 'optimal epsilon' to our test using  $\hat{\epsilon}_n = 0$  and the optimal epsilon in (6), respectively.



	B1		B2		K1		K2		K3	
const.	0.103	(0.031)	0.109	(0.03)	0.112	(0.023)	0.112	(0.021)	0.110	(0.024)
$DM_{t-1}$	0.329	(0.142)	0.350	(0.114)	0.358	(0.167)	0.363	(0.167)	0.349	(0.172)
$DM_{t-2}$	0.406	(0.098)	0.417	(0.089)	0.421	(0.174)	0.424	(0.179)	0.416	(0.175)
$FEDV_t$	0.092	(0.021)	0.097	(0.017)	0.099	(0.03)	0.100	(0.03)	0.100	(0.03)
$UN_{t-1}$	0.035	(0.01)	0.037	(0.011)	0.038	(0.008)	0.039	(0.008)	0.038	(0.008)
$R^2$	0.644		0.650		0.652		0.652		0.650	

Table 2: Estimation results for the money equation ( $\theta_1$ ) with standard errors in parentheses.

	B1		B2		K1		K2		K3	
both equations										
const.	-3.078	(0.987)	-2.864	(0.709)	-2.431	(0.123)	-2.253	(0.126)	-2.410	(0.174)
$DM_t$					-2.552	(1.412)	-2.254	(1.387)	-3.212	(3.298)
$DM_{t-1}$					-8.091	(1.069)	-6.791	(1.301)	-7.896	(1.374)
$DM_{t-2}$							-2.325	(1.003)		
$DMR_t$	-5.835	(2.185)	-4.146	(2.485)					0.862	(4.373)
$DMR_{t-1}$	-12.089	(1.903)	-11.858	(1.295)						
$DMR_{t-2}$	-4.167	(1.412)	-4.312	(1.223)						
$DG_t$					-1.068	(0.16)	-1.219	(0.149)	-1.064	(0.159)
$MIL_t$	-4.637	(0.945)	-3.638	(1.15)	-2.581	(0.618)	-3.424	(0.656)	-2.547	(0.643)
$MINW_t$	0.970	(0.861)	-0.605	(0.81)	-1.551	(0.512)	-1.764	(0.549)	-1.625	(0.761)
trend			0.014	(0.006)	0.027	(0.004)	0.025	(0.004)	0.028	(0.009)
$R^2$	0.787		0.831		0.814		0.838		0.815	
only unemployment equation										
const.	-3.078	(0.143)	-2.864	(0.14)	-2.431	(0.122)	-2.252	(0.129)	-2.410	(0.185)
$DM_t$					-2.552	(1.437)	-2.255	(1.359)	-3.212	(4.043)
$DM_{t-1}$					-8.091	(1.006)	-6.791	(1.282)	-7.896	(1.296)
$DM_{t-2}$							-2.325	(1.001)		
$DMR_t$	-5.835	(1.939)	-4.146	(1.822)					0.862	(5.093)
$DMR_{t-1}$	-12.089	(1.67)	-11.858	(1.361)						
$DMR_{t-2}$	-4.167	(1.682)	-4.312	(1.562)						
$DG_t$					-1.069	(0.159)	-1.218	(0.148)	-1.064	(0.158)
$MIL_t$	-4.637	(0.719)	-3.638	(0.698)	-2.581	(0.618)	-3.424	(0.654)	-2.547	(0.65)
$MINW_t$	0.970	(0.423)	-0.605	(0.623)	-1.551	(0.513)	-1.765	(0.557)	-1.625	(0.732)
trend			0.014	(0.004)	0.027	(0.005)	0.025	(0.004)	0.028	(0.008)
$R^2$	0.787		0.831		0.814		0.838		0.815	

Table 3: Estimation results for the unemployment equation ( $\theta_2$ ) with standard errors in parentheses.

		K1	K2	K3
both equations	B1	0.689	0.637	0.701
	B2	0.713	0.664	0.725
only unemployment equation	B1	0.627	0.613	0.692
	B2	0.658	0.649	0.722

Table 4: Estimates of the optimal  $\hat{\varepsilon}_n$ .

		K1	K2	K3
both equations	B1	-0.292	-0.854	-0.274
	B2	0.721	0.088	0.734
only unemployment equation	B1	-0.560	-1.089	-0.547
	B2	0.409	-0.232	0.421

Table 5: Value of our regularized model selection test statistic  $\tilde{t}_n$  based on the optimal  $\hat{\varepsilon}_n$ . At 5% nominal level, the test rejects when  $|\tilde{t}_n| > 1.96$ . Rejection with a positive sign of the test statistic means that the new classical model (B) is preferred over the Keynesian model (K).

	$\hat{\varepsilon}_n$		K1	K2	K3	
both equations	0.1	B1	-0.626	-1.156	-0.606	
	0.1	B2	0.487	-0.162	0.499	
	0.3	B1	-0.493	-1.037	-0.475	
	0.3	B2	0.601	-0.053	0.612	
	0.5	B1	-0.379	-0.924	-0.364	
	0.5	B2	0.673	0.033	0.684	
	0.8	B1	-0.249	-0.782	-0.237	
	0.8	B2	0.735	0.126	0.746	
	1	B1	-0.184	-0.706	-0.173	
	1	B2	0.759	0.170	0.770	
	1.2	B1	-0.132	-0.642	-0.122	
	1.2	B2	0.774	0.204	0.786	
	optimal	B1	-0.292	-0.854	-0.274	
	optimal	B2	0.721	0.088	0.734	
	only unemployment equation	0.1	B1	-0.582	-1.076	-0.583
		0.1	B2	0.418	-0.200	0.417
0.3		B1	-0.576	-1.087	-0.572	
0.3		B2	0.418	-0.215	0.422	
0.5		B1	-0.567	-1.090	-0.559	
0.5		B2	0.414	-0.226	0.423	
0.8		B1	-0.551	-1.086	-0.540	
0.8		B2	0.405	-0.237	0.419	
1		B1	-0.541	-1.080	-0.528	
1		B2	0.399	-0.242	0.416	
1.2		B1	-0.531	-1.074	-0.516	
1.2		B2	0.393	-0.246	0.412	
optimal		B1	-0.560	-1.089	-0.547	
optimal		B2	0.409	-0.232	0.421	

Table 6: Sensitivity analysis: Value of our regularized model selection test statistic  $\tilde{t}_n$  for different values of  $\hat{\varepsilon}_n$ . At 5% nominal level, the test rejects when  $|\tilde{t}_n| > 1.96$ . Rejection with a positive sign of the test statistic means that the new classical model (B) is preferred over the Keynesian model (K).

		K1	K2	K3
both equations	B1	10.319	12.047	10.479
	B2	9.160	10.464	9.310
only unemployment equation	B1	11.118	13.264	11.133
	B2	9.772	11.257	9.748

Table 7: Value of the 1st step Vuong statistic.

		K1	K2	K3
both equations	B1	29.285	30.987	29.800
	B2	34.637	33.698	33.358
only unemployment equation	B1	29.285	30.987	29.800
	B2	34.637	33.698	33.358

Table 8: 5%-level critical values for the 1st step Vuong statistic.

		K1	K2	K3
both equations	B1	-0.696	-1.212	-0.675
	B2	0.410	-0.226	0.423
only unemployment equation	B1	-0.583	-1.066	-0.586
	B2	0.416	-0.190	0.412

Table 9: Value of the 2nd step Vuong statistic. At the 5% nominal level, the test rejects when the absolute value of the test statistic is larger than 1.96.

		K1	K2	K3
both equations	B1	-0.278	-0.753	-0.094
	B2	0.576	0.172	0.744
only unemployment equation	B1	-0.496	-0.918	-0.368
	B2	0.409	-0.142	0.546

Table 10: Value of the Shi statistic.

		K1	K2	K3
both equations	B1	2.098	2.068	2.070
	B2	2.077	2.062	2.057
only unemployment equation	B1	2.003	2.060	1.996
	B2	1.972	2.047	1.996

Table 11: 5% nominal level critical values for the absolute value of the Shi statistic.