DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES

CALIFORNIA INSTITUTE OF TECHNOLOGY

PASADENA, CALIFORNIA 91125

LIKELIHOOD RATIO TESTS FOR MODEL SELECTION AND NON-NESTED HYPOTHESES

Quang H. Vuong
California Institute of Technology

SOCIAL SCIENCE WORKING PAPER 605

March 1986

LIKELIHOOD RATIO TESTS FOR

MODEL SELECTION AND NON-NESTED HYPOTHESES

Quang H. Vuong
California Institute of Technology
Division of the Humanities
and Social Sciences 228-77
Pasadena, California 91125

ABSTRACT

In this paper, we propose a classical approach to model selection. Using the Kullback-Leibler Information measure, we propose simple and directional likelihood-ratio tests for discriminating and choosing between two competing models whether the models are non-nested, overlapping or nested and whether both, one, or neither is misspecified. As a prerequisite, we fully characterize the asymptotic distribution of the likelihood ratio statistic under the most general conditions.

LIKELIHOOD RATIO TESTS FOR MODEL SELECTION AND NON-NESTED HYPOTHESES*

Quang H. Vuong
California Institute of Technology

1. INTRODUCTION

The main purpose of this paper is to propose some new tests for model selection and non-nested hypotheses. At the same time, we shall propose a classical approach to model selection. Since all our tests are based on the likelihood ratio principle, as a prerequisite, we shall completely characterize the asymptotic distribution of the likelihood ratio statistic under general conditions. By general conditions we mean that the models may be nested, non-nested or overlapping and that both, only one, or neither of the competing models may contain the true law generating the observations.

Unlike most previous work on model selection (see, e.g., Chow (1983, Chapter 9), Judge et al. (1985, Chapter 21)), we shall adopt the classical hypothesis testing framework and propose some directional and symmetric tests for choosing between models. This approach, which has not attracted a lot of attention, dates back to Hotelling (1940). A notable and recent exception is White and Olson (1979) where competing models are evaluated according to their mean square error of prediction. In this paper, we shall follow Akaike (1973, 1974) and consider the Kullback-Leibler (1951) Information Criterion (KLIC) which measures the distance between a given distribution and the true distribution. If the distance between a specified model and the true distribution is defined as the minimum of the KLIC over the distributions in the model, then it is natural to define the "best" model among a collection of competing models to be the model that is closest to the true distribution (see also Sawa (1978)).

We shall consider conditional models so as to allow for explanatory variables. Then, if $F_\theta = \{f(y|z;\theta);\ \theta \in \Theta\}$ is a conditional model, its distance from the true conditional density $h^0(y|z)$, as measured by the minimum KLIC, is $E^0[\log h^0(y|z)] - E^0[\log f(y|z;\theta_*)]$ where $E^0[\cdot]$ denotes the expectation with respect to the true joint distribution of (y,z) and θ_* is the pseudo-true value of θ (see, e.g., Sawa (1978), White (1982a)). Thus, an equivalent selection criterion can be based on the quantity $E^0[\log f(y|z;\theta_*)]$; the "best" model being the one for which this quantity is the largest.

Given two conditional models F_θ and $G_\gamma = \{g(y|z;\gamma);\ \gamma \in \Gamma\}$ which may be nested, non-nested or overlapping, we shall propose tests of the null hypothesis that $E^0[\log f(y|z;\theta_*)] = E^0[\log g(y|z;\gamma_*)]$ meaning that the two models are equivalent, against $E^0[\log f(y|z;\theta_*)] > E^0[\log g(y|z;\gamma_*)]$ meaning that F_θ is better than G_γ, or against $E^0[\log f(y|z;\theta_*)] < E^0[\log g(y|z;\gamma_*)]$ meaning that G_γ is better than F_θ. Tests of such hypotheses will be called <u>tests for model selection</u>. Since the true density $h^0_{y|z}$ is not restricted a priori to belong to either one of the parametric models F_θ and G_γ, then by necessity, the concern of this paper will solely be with asymptotic results.

The quantity $E^0[\log f(y|z;\theta_*)]$ is unknown. It can nevertheless be consistently estimated, under some regularity conditions, by (1/n) times the log-likelihood evaluated at the pseudo or quasi maximum likelihood estimator (MLE) (see e.g., White (1982a), Gourieroux, Monfort and Trognon (1984)). Hence (1/n) times the log-likelihood ratio (LR) statistic is a consistent estimator of the quantity $E^0[\log f(y|z;\theta_*)] - E^0[\log g(y|z;\gamma_*)]$. Then given the above definition of a "best" model, it is natural to consider the LR statistic as a basis for constructing tests for model selection. Since the two competing models may be nested, non-nested or overlapping, and since both, only one, or neither of the two models may be correctly specified, then it is necessary to obtain the asymptotic distribution of the LR statistic under the most general conditions. To do so, we shall use the by-now well-known framework of White (1982a) in order to handle the possibly misspecified case.

Since Neyman and Pearson (1928) advocated the LR test, it has become one of the most popular methods for testing restrictions on the parameters of a statistical model. It is well-known that minus twice the LR statistic has a limiting central chi-square distribution under the null hypothesis (Wilks (1938)), and a limiting non-central chi-square distribution under a sequence of local alternatives (Wald (1943)) with a non-centrality parameter equal to that of the Wald statistic (Wald (1943)) and Lagrange Multiplier statistic (Aitchinson and Silvey (1958), Silvey (1959)). However, as Foutz and Srivastana

(1977), Kent (1982), and White (1982a) pointed out, when the largest model is misspecified, the LR statistic is no longer necessarily chi-square distributed under the null hypothesis where the null hypothesis must be appropriately redefined in terms of the pseudo-true values satisfying the specified restrictions.

Parallel to this literature on nested hypothesis testing, the LR statistic has also been advocated as a basis for testing non-nested models (Cox (1961, 1962)). In particular Cox (1961, 1962) and White (1982b) showed that, if n denotes the sample size, then $n^{-1/2}$ times the LR statistic properly centered and normalized has a limiting standard normal distribution under the hypothesis that one of the competing models is correctly specified. This result and the result of the previous paragraph suggest that the asymptotic distribution of the LR statistic as well as the speed at which it converges to that distribution depend on whether or not the models are nested or correctly specified.

In the first part of this paper, we shall completely characterize the asymptotic distribution of the LR statistic under the most general conditions. In particular we show that the asymptotic distribution of the LR statistic and the speed at which it converges to that distribution depends on whether $f(y|z;\theta_*) = g(y|z;\gamma_*)$. In addition since the asymptotic distribution of the LR statistic depends on $f(y|z;\theta_*) = g(y|z;\gamma_*)$, we propose a test of that condition, which we call the variance test.

The paper is organized as follows. In Section 2, we present the basic framework which is that of White (1982a) and Vuong (1983, 1984). In Section 3, we derive the asymptotic distribution of the LR statistic whether or not the models are nested or misspecified. We show that: (i) if $f(y|z;\theta_*) = g(y|z;\gamma_*)$ then 2LR has a limiting weighted sum of chi-square distributions; (ii) if $f(y|z;\theta_*) \neq g(y|z;\gamma_*)$ then $n^{-1/2}$ LR properly centered around $E^0[\log(f(y|z;\theta_*)/g(y|z;\gamma_*))]$ has a limiting normal distribution with non-zero variance ω_*^2. In addition, for the first case, we characterize the conditions under which 2LR is asymptotically chi-square distributed.

In Section 4, we show that $f(y|z;\theta_*) = g(y|z;\gamma_*)$ is equivalent to the hypothesis that the variance $\omega_*^2 = 0$. This allows us to construct a test of the hypothesis $f(y|z;\theta_*) = g(y|z;\gamma_*)$ based on a consistent estimator $\hat{\omega}_n^2$ of ω_*^2. We show that $n\hat{\omega}_n^2$ has a limiting weighted sum of chi-square distributions under the null hypothesis $\omega_*^2 = 0$ and we also characterize the cases for which this limiting distribution reduces to a chi-square distribution.

In the next three sections, we apply the previous results to derive LR based tests for model selection in all possible situations. The case where the models are (strictly) non-nested is considered in Section 5. There, we propose a new and very simple directional test based on $n^{-1/2}LR/\hat{\omega}_n$, for selecting the best of two models. The statistic has a limiting standard normal distribution under the null hypothesis that the two non-nested models are equivalent, whether or not both, one or neither is misspecified. We also discuss the

relationship between our approach to model selection and that of Akaike (1973, 1974).

In Section 6, we consider the case where the models are overlapping. This case is seen to be more complicated than the nested case since, under the null hypothesis that the models are equivalent, the asymptotic distribution of the LR statistic depends on whether or not $\omega_*^2 = 0$. We propose two procedures. The first procedure is used when ω_*^2 is possibly null under the null hypothesis that the models are equivalent. The procedure is sequential and is based on the variance statistic of Section 4 for testing $\omega_*^2 = 0$ followed by the normal LR test of Section 5 in case of rejection of $\omega_*^2 = 0$. The second procedure applies when ω_*^2 is always null under the null hypothesis that the models are equivalent. This happens, as we shall show, when one of the two overlapping models is correctly specified. Then a model selection test can be based directly on twice the LR statistic.

Finally Section 7 considers the more familiar case of nested models. We show that testing restrictions on θ_* is actually identical to testing that the two models are equivalent against the hypothesis that the largest model is "best." Thus, when the competing models are nested, our model selection approach coincides with the classical hypothesis testing approach. Then we propose a test based on twice the LR statistic which reduces to the familiar Neyman-Pearson LR test when for instance the largest model is correctly specified. We also propose a new test based on the variance statistic of Section 4 for testing restrictions on θ_* which can also be interpreted as a model selection test.

Section 8 summarizes our results, suggests some directions for further research, and contains our view on the general purpose of model selection and hypothesis testing in econometric modelling. In particular, we discuss the important distinction between our tests for model selection and the non-nested hypotheses tests proposed by Cox (1961, 1962). All the proofs are collected in the Appendix.

2. BASIC FRAMEWORK

Let X_t be a $m \times 1$ observed random vector defined on an Euclidean measurable space (X, σ, ν_x). For instance, in the case of a continuous random vector, X, σ, ν_x are respectively $\mathbb{R}^m$, the Borel σ-algebra, and the usual Lebesgue measure. The process generating the observation X_t, $t = 1,2,\ldots$ satisfies the following assumption.

<u>Assumption A1</u>: The random vectors X_t, $t = 1,2,\ldots$ are independent and identically distributed (i.i.d.) with common true cumulative distribution function H^0 on (X, σ, ν_x).

Though there are more general assumptions on the true data generating process than Assumption A1 (see, e.g., Gallant and Holly (1980), Burguette, Gallant and Souza (1982)), Assumption A1 is the simplest assumption that still allows for the presence of exogenous variables. Following Vuong (1983), the vector X_t is partitioned into $X_t = (Y_t', Z_t')'$ where Y_t and Z_t are respectively ℓ and k dimensional vectors with $m = \ell + k$. Let (Y, σ_y, ν_y) and (Z, σ_z, ν_z) be the Euclidean measurable spaces associated with Y_t and Z_t. We shall be interested in the true conditional distribution $H^0_{Y|Z}(\cdot|\cdot)$ of Y_t given Z_t. It is convenient to think of Y_t as being the endogenous variables, and of Z_t as being the exogenous variables.

We now consider two competing parametric families of conditional distributions for Y_t given Z_t:
$F_\theta \equiv \{F_{Y|Z}(\cdot|\cdot;\theta);\ \theta \in \Theta \subset \mathbb{R}^p\}$ and $G_\gamma \equiv \{G_{Y|Z}(\cdot|\cdot;\gamma);\ \gamma \in \Gamma \subset \mathbb{R}^q\}$.
No assumption is here made on the relationship between the two competing conditional models F_θ and G_γ in the sense that they may be nested, overlapping, or non-nested. Moreover, both, only one, or neither may be correctly specified, i.e., may contain the true conditional distribution for Y_t given Z_t. Each conditional model satisfies, however, the following regularity conditions (Vuong (1983)) which are similar to those of White (1982, Assumptions A2-A6) with the exception that they bear on conditional models. These regularity conditions are presented without discussion. They are stated in terms of the conditional model F_θ. It is understood that similar assumptions are made on the conditional model G_γ.

<u>Assumption A2</u>: (a) Θ is a compact subset of $\mathbb{R}^k$, and for every θ in Θ and for all z the conditional distribution $F_{Y|Z}(\cdot|z;\theta)$ has a density with respect to ν_y: $f(\cdot|z;\theta) = dF_{Y|Z}(\cdot|z;\theta)/d\nu_y$. (b) The conditional density $f(y|z;\theta)$ is a strictly positive function that is measurable in (y,z) for any θ, and continuous in θ for all (y,z).

<u>Assumption A3</u>: (a) For (H^0-almost) all (y,z), $|\log f(y|z;\cdot)|$ is dominated by an H^0-integrable function independent of θ. (b) The function $z_f(\theta) = \int \log f(y|z;\theta)dH^0(x)$ has a unique maximum on Θ at θ_*.

The value θ_* is called the pseudo-true value of θ for the conditional model F_θ (see, e.g., Sawa (1978)). Similarly γ_* denotes the pseudo-true value of γ for the conditional model G_γ.

<u>Assumption A4</u>: (a) For (H^0-almost) all (y,z), $\log f(y|z;\cdot)$ is twice continuously differentiable on Θ. (b) For (H^0-almost) all (y,z), $|\partial \log f(y|z;\theta)/\partial\theta \cdot \partial \log f(y|z;\theta)/\partial\theta'|$ and $|\partial^2 \log f(y|z;\theta)/\partial\theta\partial\theta'|$ are dominated by H^0-integrable functions independent of θ.

This ensures the existence of the usual matrices:

$$A_f(\theta) = E^0\left[\frac{\partial^2 \log f(Y_t|Z_t;\theta)}{\partial\theta\partial\theta'}\right], \tag{2.1}$$

$$B_f(\theta) = E^0\left[\frac{\partial \log f(Y_t|Z_t;\theta)}{\partial\theta} \cdot \frac{\partial \log f(Y_t|Z_t;\theta)}{\partial\theta'}\right], \tag{2.2}$$

where $E^0[\cdot]$ denotes the expectation with respect to the true joint distribution of $X_t = (Y_t,Z_t)$. Similar matrices $A_g(\gamma)$ and $B_g(\gamma)$ are defined for the conditional model G_γ.

<u>Assumption 5</u>: (a) θ_* is an interior point of Θ. (b) θ_* is a regular point of $A_f(\theta)$.

Assumptions A1-A5 can be thought of as the simplest regularity assumptions for maximum likelihood estimation under general conditions in the presence of explanatory variables. The (quasi) maximum likelihood (ML) estimator $\hat{\theta}_n$ for the conditional model F_θ is a σ_X^n-measurable function of $(X_1,\ldots,X_n)$ such that

$$L_n^f(\hat{\theta}_n) = \sup_{\theta\varepsilon\Theta} L_n^f(\theta), \tag{2.3}$$

where $L_n^f(\theta)$ is the (conditional) log-likelihood function for the model F_θ:

$$L_n^f(\theta) \equiv \sum_{t=1}^{n} \log f(Y_t|Z_t;\theta). \tag{2.4}$$

A similar definition applies to the ML estimator $\hat{\gamma}_n$ for the conditional model G_γ with respect to the log-likelihood function:

$$L_n^g(\gamma) \equiv \sum_{t=1}^{n} \log g(Y_t|Z_t;\gamma). \tag{2.5}$$

Given Assumptions A1-A5, it follows from White (1982a) among others that the ML estimator $\hat{\theta}_n$ exists, is consistent for θ_*, and is asymptotically normally distributed with asymptotic covariance matrix $A_f^{-1}(\theta_*)B_f(\theta_*)A_f^{-1}(\theta_*)$. Moreover the asymptotic covariance matrix can be consistently estimated by $A_{fn}^{-1}(\hat{\theta}_n)B_{fn}(\hat{\theta}_n)A_{fn}^{-1}(\hat{\theta}_n)$ where $A_{fn}(\theta)$ and $B_{fn}(\theta)$ are the sample analogs of $A_f(\theta)$ and $B_f(\theta)$. That is:

$$A_{fn}(\theta) = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial^2 \log f(Y_t|Z_t;\theta)}{\partial\theta\partial\theta'}, \tag{2.6}$$

$$B_{fn}(\theta) = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial \log f(Y_t|Z_t;\theta)}{\partial\theta} \cdot \frac{\partial \log f(Y_t|Z_t;\theta)}{\partial\theta'}. \tag{2.7}$$

Similar properties hold for the ML estimator $\hat{\gamma}_n$ of γ_*.

In the next section, we shall need the joint asymptotic distribution of $\hat{\theta}_n$ and $\hat{\gamma}_n$. Since A4 holds for both models F_θ and G_γ, then it can be shown that for (H^0-almost) all (y,z), $|\partial \log f(y|z;\cdot)/\partial\theta \cdot \partial \log g(y|z;\cdot)/\partial\gamma'|$ is dominated by an H^0-integrable function independent of θ and γ. This ensures that the $p \times q$ matrix

$$B_{fg}(\theta,\gamma) = B'_{gf}(\gamma,\theta) \equiv E^0\left[\frac{\partial \log f(Y_t|Z_t;\theta)}{\partial\theta} \cdot \frac{\partial \log g(Y_t|Z_t;\gamma)}{\partial\gamma'}\right] \tag{2.8}$$

exists. Moreover, from Jennrich's uniform strong Law of Large Numbers (1969, Theorem 2), it follows that $B_{fg}(\theta_*,\gamma_*)$ is consistently estimated by its sample analog:

$$B_{fgn}(\hat{\theta}_n,\hat{\gamma}_n) = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial \log f(Y_t|Z_t;\hat{\theta}_n)}{\partial\theta} \cdot \frac{\partial \log g(Y_t|Z_t;\hat{\gamma}_n)}{\partial\gamma'}. \tag{2.9}$$

The next lemma gives the joint asymptotic distribution for the quasi ML estimators $\hat{\theta}_n$ and $\hat{\gamma}_n$.

<u>Lemma</u> 2.1: Given Assumptions A1-A5:

$$n^{1/2}\begin{bmatrix} \hat{\theta}_n - \theta_* \\ \hat{\gamma}_n - \gamma_* \end{bmatrix} \xrightarrow{D} N(0,\Sigma)$$

where

$$\Sigma = \begin{bmatrix} A_f^{-1}(\theta_*)B_f(\theta_*)A_f^{-1}(\theta_*) & ; & A_f^{-1}(\theta_*)B_{fg}(\theta_*,\gamma_*)A_g^{-1}(\gamma_*) \\ A_g^{-1}(\theta_*)B_{gf}(\gamma_*,\theta_*)A_f^{-1}(\gamma_*) & ; & A_g^{-1}(\gamma_*)B_g(\gamma_*)A_g^{-1}(\gamma_*) \end{bmatrix}. \tag{2.10}$$

Moreover, the asymptotic covariance matrix Σ can be consistently estimated by $\hat{\Sigma}_n$ which is defined as in Equation (2.10) where A and B are replaced by their sample analogs evaluated at the ML estimators $\hat{\theta}_n$ and $\hat{\gamma}_n$.

3. THE LIKELIHOOD RATIO STATISTIC

All the tests for model selection that are proposed later in this paper will be based on the likelihood ratio (LR) statistic. In this section, we shall therefore obtain the asymptotic distribution of the LR statistic under the most general conditions.

The LR statistic for the model F_θ against the model G_γ is defined as:

$$LR_n(\hat{\theta}_n,\hat{\gamma}_n) \equiv L_n^f(\hat{\theta}_n) - L_n^g(\hat{\gamma}_n)$$

$$= \sum_{t=1}^{n} \log \frac{f(Y_t|Z_t;\hat{\theta}_n)}{g(Y_t|Z_t;\hat{\gamma}_n)} \tag{3.1}$$

where $\hat{\theta}_n$ and $\hat{\gamma}_n$ are the ML estimators of θ_* and γ_* defined in the previous section.

<u>Lemma</u> 3.1: Given Assumptions A1-A3:

$$\frac{1}{n}LR_n(\hat{\theta}_n,\hat{\gamma}_n) \xrightarrow{a.s.} E^0\left[\log \frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right]. \tag{3.2}$$

This result is important because it motivates our LR-based tests for model selection. To derive the asymptotic distribution of the LR statistic, we use the following lemma.

<u>Lemma 3.2</u>: Given Assumptions A1-A5:

(i) if $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$, then:

$$LR_n(\hat{\theta}_n,\hat{\gamma}_n) = -\frac{n}{2}(\hat{\theta}_n - \theta_*)' A_f(\theta_*)(\hat{\theta}_n - \theta_*) + \frac{n}{2}(\hat{\gamma}_n - \gamma_*)' A_g(\gamma_*)(\hat{\gamma}_n - \gamma_*) + o_p(1), \tag{3.3}$$

(ii) if $f(\cdot|\cdot;\theta_*) \neq g(\cdot|\cdot;\gamma_*)$, then:

$$LR_n(\hat{\theta}_n,\hat{\gamma}_n) = LR_n(\theta_*,\gamma_*) + O_p(1).^1 \tag{3.4}$$

The condition $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ is to be understood as meaning that $f(y|z;\theta_*) = g(y|z;\gamma_*)$ for H^0-almost all (y,z). Lemma 3.2 shows that the asymptotic distribution of the LR statistic depends on whether or not $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$. This latter condition will be considered in the next section. Let us note that if the two models F_θ and G_γ are strictly non-nested, as defined later, then one must have $f(\cdot|\cdot;\theta_*) \neq g(\cdot|\cdot;\gamma_*)$. On the other hand, if the models F_θ and G_γ are nested or overlapping, then one may have $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$.

If this latter condition holds, then the first part of Lemma 3.2 states that the LR statistic is asymptotically distributed as a quadratic form in $n^{1/2}(\hat{\theta}_n - \theta_*)$ and $n^{1/2}(\hat{\gamma}_n - \gamma_*)$ which are asymptotically normal as shown in Lemma 2.1. It is therefore important to consider the distributions of quadratic forms in normal

random variables. Such distributions have already been studied (see, e.g. Johnson and Kotz (1970, Chapter 29)). We call such distributions, weighted sums of (independent) chi-square distributions, for which we give the following definition.

<u>Definition 3.3</u> (Weighted Sums of Chi-Square Distributions): Let $Z = (Z_1,\ldots,Z_m)'$ be a vector of m independent standard normal variables, and let $\lambda = (\lambda_1,\ldots,\lambda_m)'$ be a vector of m real numbers. Then, the random variable

$$Q(Z) = \sum_{i=1}^{m} \lambda_i Z_i^2 \tag{3.5}$$

is distributed as a weighted sum of chi-square distributions with parameters (m,λ). Its cumulative distribution function (c.d.f.) is denoted by $M_m(\cdot;\lambda)$.

Let us note that the distribution of $Q(Z)$ depends only on the non-zero parameters λ_i. In other words, the c.d.f. $M_m(\cdot;\lambda)$ is identically equal to the c.d.f. $M_{\underline{m}}(\cdot;\underline{\lambda})$ where $\underline{\lambda}$ is the vector of non-zero λ_i's, and $\underline{m}$ is the number of such λ_i's. Moreover, the mixture $M_m(\cdot;\lambda)$ reduces to a central chi-square distribution if and only if the non-zero parameters λ_i are equal to one, in which case the number of degrees of freedom is equal to $\underline{m}$.

The next lemma shows that any quadratic form in m random variables that are jointly normally distributed with zero means and some covariance matrix Ω is distributed as a weighted sum of chi-squares with some parameters m and λ. This results allows Ω to be singular, and slightly differs from Moore (1978, Theorem 1).

<u>Lemma 3.4</u>: Let Y be a vector of m random variables distributed as $N(0,\Omega)$ with rank $\Omega \equiv r \leq m$. Let Q be a $m \times m$ real symmetric matrix. Then the quadratic form

$$Q(Y) \equiv Y'QY \sim M_m(\cdot;\lambda) \tag{3.6}$$

where λ is the vector of eigenvalues of $Q\Omega$.[2]

We can now readily obtain the asymptotic distribution of the LR statistic under general conditions. Let ω_*^2 denote the variance of $\log[f(Y_t|Z_t;\theta_*)/g(Y_t|Z_t;\gamma_*)]$ where the variance is computed with respect to the true joint distribution H^0 of (Y_t,Z_t). That is:

$$\omega_*^2 \equiv \mathrm{var}^0\left[\log \frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right]$$

$$= E^0\left[\log \frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right]^2 - \left[E^0\left[\log \frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma^*)}\right]\right]^2 . \tag{3.7}$$

To ensure that such a variance exists, we make the following assumption.

<u>Assumption A6</u>: For (H^0-almost) all (y,z) the functions $|\log f(y|z;\cdot)|^2$ and $|\log g(y|z;\cdot)|^2$ are dominated by H^0-integrable functions independent of θ and γ.

<u>Theorem 3.5</u> (Asymptotic Distribution of the LR Statistic): Given Assumption A1-A6:

(i) if $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$, then

$$2LR_n(\hat\theta_n,\hat\gamma_n) \xrightarrow{D} M_{p+q}(\cdot;\lambda_*), \tag{3.8}$$

where λ_* is the vector of p + q eigenvalues of

$$W = \begin{bmatrix} -B_f(\theta_*)A_f^{-1}(\theta_*) & ; & -B_{fg}(\theta_*,\gamma_*)A_g^{-1}(\gamma_*) \\ B_{gf}(\gamma_*,\theta_*)A_f^{-1}(\theta_*) & ; & B_g(\gamma_*)A_g^{-1}(\gamma_*) \end{bmatrix},^{3} \tag{3.9}$$

(ii) if $f(\cdot|\cdot;\theta_*) \neq g(\cdot|\cdot;\gamma_*)$, then

$$n^{-1/2}LR_n(\hat\theta_n,\hat\gamma_n) - n^{1/2}E^0\left[\log \frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] \xrightarrow{D} N(0,\omega_*^2).^{4} \tag{3.10}$$

Theorem 3.5 characterizes the asymptotic distribution of the LR statistic under general conditions. It shows that the asymptotic distribution of the LR statistic as well as the speed at which it converges to that distribution depends on whether or not $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$.

The limiting weighted sum of chi-square distributions that arises when $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ is somewhat unusual. It is therefore useful to characterize the conditions under which this limiting distribution reduces to the familiar chi-square distribution. This is the purpose of the next result. For this result, we shall however assume that the information matrix equivalence holds for both conditional models F_θ and G_γ, i.e.:

$$A_f(\theta_*) + B_f(\theta_*) = 0 \text{ and } A_g(\gamma_*) + B_g(\gamma_*) = 0. \tag{3.11}$$

As mentioned in White (1982a, Theorem 3.3) and Vuong (1983, Lemma 3), the information matrix equivalences hold under correct specification of the conditional models given mild additional assumptions.

Theorem 3.6 (Asymptotic Chi-Square Distribution of the LR Statistic given Information Matrix Equivalences): Given Assumptions A1–A5, suppose that Equation (3.11) holds. If $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$, then $2LR_n(\hat{\theta}_n,\hat{\gamma}_n)$ converges to a central chi-square distribution if and only if:

$$B_g(\gamma_*) - B_{gf}(\gamma_*,\theta_*)B_f^{-1}(\theta_*)B_{fg}(\theta_*,\gamma_*) = 0, \tag{3.12}$$

in which case the number of degrees of freedom is $p - q$.

As seen in Section 7 below, Condition (3.12) will be satisfied when the conditional model G_γ is nested in the conditional model F_θ.

4. THE VARIANCE STATISTIC

In the previous section, we showed that whether the LR statistic is asymptotically distributed as a normal or a weighted sum of chi-squares depends on whether or not $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$. As mentioned there, this latter equality may hold when the conditional models F_θ and G_γ are nested or overlapping. It is therefore important to know if such a condition is satisfied. Since θ_* and γ_* are unknown, we shall propose in this section a test of such a condition. The proposed test is based on the following property.

Lemma 4.1: Given Assumptions A2, A3, and A6, $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ if and only if $\omega_*^2 = 0$.

The importance of Lemma 4.1 is that to test the crucial condition $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ one can equivalently test the condition that the variance ω_*^2 is equal to zero. We define the following null and alternative hypotheses:

$$H_0^\omega:\ \omega_*^2 = 0 \quad \text{vs.} \quad H_A^\omega:\ \omega_*^2 \neq 0. \tag{4.1}$$

Then a natural statistic that we can use to test H_0^ω against H_A^ω is the sample analog:

$$\hat{\omega}_n^2 \equiv \frac{1}{n}\sum_{t=1}^{n}\left[\log\frac{f(Y_t|Z_t;\hat{\theta}_n)}{g(Y_t|Z_t;\hat{\gamma}_n)}\right]^2 - \left[\frac{1}{n}\sum_{t=1}^{n}\log\frac{f(Y_t|Z_t;\hat{\theta}_n)}{g(Y_t|Z_t;\hat{\gamma}_n)}\right]^2. \tag{4.2}$$

Moreover, let us note that ω_*^2 is also the variance of the limiting normal distribution of the LR statistic (see Theorem 3.4 – (ii)). Thus the variance statistic $\hat{\omega}_n^2$ will play two important roles: first, to be a basis for a test of $\omega_*^2 = 0$ or equivalently $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$; second, to be an estimator of the asymptotic variance of the LR statistic when $f(\cdot|\cdot;\theta_*) \neq g(\cdot|\cdot;\gamma_*)$.

An alternative variance statistic that will play a similar role and that is even easier to compute than $\hat{\omega}_n^2$ is:

$$\tilde{\omega}_n^2 \equiv \frac{1}{n}\sum_{t=1}^{n}\left[\log\frac{f(Y_t|Z_t;\hat{\theta}_n)}{g(Y_t|Z_t;\hat{\gamma}_n)}\right]^2. \tag{4.3}$$

Note that from Equations (3.1) and (4.2), we have:

$$\tilde{\omega}_n^2 = \hat{\omega}_n^2 + (\tfrac{1}{n}LR_n(\hat{\theta}_n,\hat{\gamma}_n))^2 \geq \hat{\omega}_n^2. \tag{4.4}$$

The next lemma states that these variance statistics are strongly consistent estimators of their population analogs.

Lemma 4.2: Given Assumptions A1-A3, and A6:

(i) $$\hat{\omega}_n^2 \overset{a.s.}{\to} \omega_*^2, \tag{4.5}$$

(ii) $$\tilde{\omega}_n^2 \overset{a.s.}{\to} \omega_*^2 + \left[E^0\left[\log \frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right]\right]^2. \tag{4.6}$$

To construct a test of H_0^ω against H_A^ω, it is necessary to derive the asymptotic distribution of the variance statistic $\hat{\omega}_n^2$ or $\tilde{\omega}_n^2$. We make the following assumption.

Assumption A7: For (H^0-almost) all (y,z) the functions $|\log[f(y|z;\cdot)/g(y|z;\cdot)] \cdot \partial^2\log f(y|z;\cdot)/\partial\theta\partial\theta'|$ and $|\log[f(y|z;\cdot)/g(y|z;\cdot)] \cdot \partial^2\log g(y|z;\cdot)/\partial\gamma\partial\gamma'|$ are dominated by H^0-integrable functions idependent of θ and γ.

Theorem 4.3 (Asymptotic Distribution of the Variance Statistics given $\omega_*^2 = 0$): Given Assumptions A1-A7, under H_0^ω: $\omega_*^2 = 0$, we have:

$$n\hat{\omega}_n^2 = n\tilde{\omega}_n^2 + o_p(1) \overset{D}{\to} M_{p+q}(\cdot,\lambda_*^2) \tag{4.7}$$

where λ_*^2 is the vector of squares of the p + q eigenvalues λ_* of W.

Theorem 4.3 says that, under the null hypotheses H_0^ω, the two statistics $n\hat{\omega}_n^2$ and $n\tilde{\omega}_n^2$ are asymptotically equivalent, and have a limiting distribution which is again a weighted sum of chi-squares. The parameter λ_*^2 are, as expected, all non-negative. This contrasts with the parameters λ_* of the limiting weighted sum of chi-squares for the LR statistic which may be negative (see footnote 3).

As for the LR statistic, it is of interest to know when the limiting wieghted sum of chi-squares distribution of the variance statistics reduce to the familiar central chi-square distribution. The next result characterizes this situation. As for Theorem 3.6, we assume that the information matrix equivalences (3.11) hold.

Theorem 4.4 (Asymptotic Chi-Square Distribution of the Variance Statistics given Information Matrix Equivalences and $\omega_*^2 = 0$): Given Assumptions A1-A7, suppose that Equation (3.11) holds. Then, under H_0^ω: $\omega_*^2 = 0$, the following are equivalent:

(i) $n\hat{\omega}_n^2$ converges in distribution to a chi-square,

(ii) $n\tilde{\omega}_n^2$ converges in distribution to a chi-square,

(iii) $B_{fg}(\theta_*,\gamma_*)B_g^{-1}(\gamma_*)B_{gf}(\gamma_*,\theta_*)B_f^{-1}(\theta_*)$ is idempotent,

(iv) $B_{gf}(\gamma_*,\theta_*)B_f^{-1}(\theta_*)B_{fg}(\theta_*,\gamma_*)B_g^{-1}(\gamma_*)$ is idempotent,

in which case the number of degrees of freedom is p + q − 2 rank $B_{fg}(\theta_*,\gamma_*)$.

As shown in Section 7 below, conditions (iii) or (iv) will be satisfied if G_γ is nested in F_θ or if F_θ is nested in G_γ. Conditions (iii) or (iv) can, however, be satisfied even when the models are non-nested or overlapping. In particular, it is easy to see that these conditions are satisfied when the conditional models F_θ and G_γ are asymptotically orthogonal as defined by Gourieroux, Monfort and Trognon (1983), i.e., when:

$$B_{fg}(\theta_*,\gamma_*) = 0, \qquad (4.8)$$

in which case the number of degrees of freedom of the limiting chi-square distribution of $n\hat{\omega}_n^2$ or $n\tilde{\omega}_n^2$ is p + q.

5. STRICTLY NON-NESTED MODELS

In section 1, we suggested a classical approach for selecting among competing models. In this section, we shall discuss this approach in more detail. In particular, using the results of Section 3 and 4, we shall obtain a very simple test for selecting among two non-nested models. Then we shall discuss the fundamental differences between our model selection approach and the more familiar one introduced by Akaike (1973, 1974).

Following Akaike (1973, 1974), Sawa (1978) and Chow (1981), our approach is based on the minimum KLIC which measures the distance between the true distribution and a specified model. For a conditional model F_θ, this measure gives:

$$KLIC(H^0_{Y|Z};F_\theta) \equiv E^0[\log h^0(Y_t|Z_t)] - E^0[\log f(Y_t|Z_t;\theta_*)], \qquad (5.1)$$

where $h^0(\cdot|\cdot)$ is the true conditional density of Y_t given Z_t, and θ_* are the pseudo-true values of θ defined in Assumption 3.[5] From

Jensen's inequality, the measure (5.1) is always non-negative and is equal to zero if and only if $h^0(\cdot|\cdot) = f(\cdot|\cdot;\theta_*)$ H^0-almost surely, i.e., if and only if the conditional model F_θ is correctly specified. Moreover, since the first term in the right-hand side of Equation (5.1) does not depend on the conditional model F_θ, then an equivalent measure is $E^0[\log f(Y_t|Z_t;\theta_*)]$.

Given a collection of competing conditional models, it is natural to select the model that is closest to the true conditional distribution. Given the above measure of distance, we shall consider the following hypotheses and definitions:

$$H_0:\ E^0\left[\log \frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] = 0, \qquad (5.2)$$

meaning that F_θ and G_γ are <u>equivalent</u>, against

$$H_f:\ E^0\left[\log \frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] > 0, \qquad (5.3)$$

meaning that F_θ is <u>better</u> than G_γ, or

$$H_g:\ E^0\left[\log \frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] < 0, \qquad (5.4)$$

meaning that F_θ is <u>worse</u> that G_γ. Tests of H_0 against H_f or H_g will be called tests for model selection. There are, of course, alternative definitions, some of which will be discussed later in this section.

The indicator $E^0[\log f(Y_t|Z_t;\theta_*)] - E^0[\log g(Y_t|Z_t;\gamma_*)]$ is unknown since θ_*, γ_*, and the joint distribution H^0 of (Y_t,Z_t) with respect to which the expectation $E^0[\cdot]$ is evaluated are all unknown. But it is clear that we can consistently estimate this unknown indicator by (1/n) times the LR statistic (see Lemma 3.1). Thus the LR statistic is a natural statistic for discriminating between two models.

In this section, we shall consider the case where the models F_θ and G_γ are (strictly) non-nested. Since Cox (1961, 1962) initial work, non-nested models have attracted a lot of interest from econometricians (see, e.g., Mackinnon (1983) recent survey and the special issue of the Journal of Econometrics edited by White (1983)). We shall first give a formal definition of strictly non-nested models.

<u>Definition 5.1</u> (Strictly Non-Nested Models): Two conditional models F_θ and G_γ are strictly non-nested if and only if:

$$F_\theta \cap G_\gamma = \emptyset.^6 \tag{5.5}$$

For instance, this is the case when F_θ and G_γ are standard linear regression models with different distributional assumptions on the errors, say normally or logistic distributed. Alternatively, the competing regressions models may have the same distributional assumption on the errors but different functional forms such as the linear or the exponential form.

Since the conditional models F_θ and G_γ do not have any conditional distribution in common, it must be the case that $f(\cdot|\cdot;\theta_*) \neq g(\cdot|\cdot;\gamma_*)$. It follows that the second part of Theorem 3.5 applies. Moreover, from Lemma 4.2, the asymptotic variance ω_*^2 can be

consistently estimated by $\hat{\omega}_n^2$ or by $\tilde{\omega}_n^2$ under the null hypothesis that the models F_θ and G_γ are equivalent, i.e., under H_0. Thus we have the following straightforward model selection test. Let $\hat{\omega}_n$ and $\tilde{\omega}_n$ be the positive square roots of $\hat{\omega}_n^2$ and $\tilde{\omega}_n^2$ respectively.

<u>Theorem 5.2</u> (Model Selection Tests for Strictly Non-Nested Models): Given Assumptions A1-A6,

(i) under H_0: $$n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n \xrightarrow{D} N(0,1), \tag{5.6}$$

(ii) under H_f: $$n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n \xrightarrow{a.s.} +\infty, \tag{5.7}$$

(iii) under H_g: $$n^{-1/2}LR_n(\hat{\theta}_n,\tilde{\gamma}_n)/\hat{\omega}_n \xrightarrow{a.s.} -\infty, \tag{5.8}$$

(iv) properties (i)-(iii) hold if $\hat{\omega}_n$ is replaced by $\tilde{\omega}_n$.

Theorem 5.2 provides a very simple directional test for model selection. Specifically, one chooses a critical value c from the standard normal distribution for some significance level. If the value of the statistic $n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n$ is higher than c then one rejects the null hypothesis that the models are equivalent in favor of F_θ being better than G_γ. If $n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n$ is smaller than $-c$ then one rejects the null hypothesis in favor of G_γ being better than F_θ. Finally if $|n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n| \leq c$ then one cannot discriminate between the two competing models given the data. Similar inferences can of course be made based on the other statistic

$n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\tilde{\omega}_n$.[7]

Let us note that these statistics are extremely easy to compute. Indeed from Equations (3.1), (4.2) and (4.3) these statistics are:

$$\frac{n^{-1/2}LR_n(\hat{\theta},\hat{\gamma})}{\hat{\omega}_n} = \frac{LR_n(\hat{\theta}_n,\hat{\gamma}_n)}{\left\{\sum_{t=1}^{n}\left[\log\frac{f(Y_t|Z_t;\hat{\theta}_n)}{g(Y_t|Z_t;\hat{\gamma}_n)}\right]^2 - \frac{1}{n}\left[LR_n(\hat{\theta}_n,\hat{\gamma}_n)\right]^2\right\}^{1/2}}, \qquad (5.9)$$

$$\frac{n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)}{\tilde{\omega}_n} = \frac{LR_n(\hat{\theta}_n,\hat{\gamma}_n)}{\left\{\sum_{t=1}^{n}\left[\log\frac{f(Y_t|Z_t;\hat{\theta}_n)}{g(Y_t|Z_t;\hat{\gamma}_n)}\right]^2\right\}^{1/2}}. \qquad (5.10)$$

Hence both statistics are equal to the difference in the maximum log-likelihood values for the two models suitably normalized. The normalization in Equation (5.10) is directly obtained from the sum of squares of $m_t = \log[f(Y_t|Z_t;\hat{\theta}_n)/g(Y_t|Z_t;\hat{\gamma}_n)]$, while the normalization in Equation (5.9) is obtained from the sum of squared deviations of m_t from its sample mean which is equal to $\frac{1}{n}LR_n(\hat{\theta}_n,\hat{\gamma}_n)$. Alternatively, these statistics can be readily obtained from an additional linear regression. For instance, it can be shown that the statistic (5.9) is numerically equal to $[(n-1)/n]^{1/2}$ times either the usual t-statistic on the constant term in a linear regression of m_t on only the constant term, or the usual t- statistic on the coefficient of m_t in a linear regression of 1 on m_t.[8]

We now contrast our approach to the more familiar approach initiated by Akaike (1973, 1974). First, as in the model selection literature, our statistics (5.9) and (5.10) can be thought of as defining a criterion for selecting among competing models. Omitting the normalizing factors $\hat{\omega}_n$ and $\tilde{\omega}_n$, our criterion is based on the <u>uncorrected</u> log-likelihood $L_n^f(\hat{\theta}_n)$ of a model. Thus to decide which model is "best" one can directly compare the maximum values of the log-likelihoods of the competing models and choose the model with the highest log-likelihood. Our criterion is very intuitive. It contrasts with the previous model selection criteria that are based on the maximum log-likelihood <u>corrected</u> for the number of estimated parameters (Akaike (1973, 1974), Sawa (1978), Schwarz (1978), Chow (1981)). Such a difference arises for the reason that these latter model selection criteria were initially derived, not as estimates of $E^0[\log f(Y_t,Z_t;\theta_*)]$, but as approximations to the alternative criterion $E_{\hat{\theta}_n}[E^0[\log f(Y_t|Z_t;\hat{\theta}_n)]]$ where $E_{\hat{\theta}_n}[\cdot]$ is the expectation with respect to the (asymptotic) distribution of the ML estimator $\hat{\theta}_n$, and $E^0[\cdot]$ is the expectation with respect to the true (joint) distribution of (Y_t,Z_t) where $\hat{\theta}_n$ is treated as a constant (see Sawa (1978, Rule 2.1 - (ii)), and Chow (1981)).[9]

Lien and Vuong (1986) pointed out, however, that each of these well- known model selection criteria can be thought of as a consistent estimate of $E^0[\log f(Y_t|Z_t;\theta_*)]$. In addition, each of these model selection criterion appropriately normalized is asymptotically equivalent to the LR-statistics (5.9) and (5.10) under the null hypothesis that the models are (KLIC) equivalent, i.e., under H_o. More generally, let

$$\tilde{LR}_n(\hat{\theta}_n,\hat{\gamma}_n) \equiv LR_n(\hat{\theta}_n,\hat{\gamma}_n) - K_n(F_\theta,G_\gamma) \quad (5.11)$$

where $K_n(F_\theta,G_\gamma)$ is a correction factor depending on the characteristics of the competing models F_θ and G_γ. We have:

Corollary 5.3 (Equivalent Model Selection Tests of Strictly Non-Nested Models): Given Assumptions A1-A6, suppose that

$$n^{-1/2}K_n(F_\theta,G_\gamma) = o_p(1). \quad (5.12)$$

(i) under H_0: $n^{-1/2}\tilde{LR}_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n \xrightarrow{D} N(0,1)$,

(ii) under H_f: $n^{-1/2}\tilde{LR}_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n \xrightarrow{a.s.} +\infty$,

(iii) under H_g: $n^{-1/2}\tilde{LR}_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n \xrightarrow{a.s.} -\infty$.

This result follows by noticing that:

$$n^{-1/2}\tilde{LR}_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n = n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n + o_p(1). \quad (5.13)$$

It also follows that $\tilde{\omega}_n$ can equivalently replace $\hat{\omega}_n$ in Corollary 5.3. Example of correction factors that satisfy (5.12) are $K_n(F_\theta,G_\gamma) = p - q$ and $K_n(F_\theta,G_\gamma) = \frac{p}{2}\log n - \frac{q}{2}\log n$, which correspond to Akaike (1973) and Schwarz (1978) information criteria.

Corollary 5.3 implies that one can also use the corrected log-

likelihood ratio $\tilde{LR}_n(\hat{\theta}_n,\hat{\gamma}_n)$ as a basis for a model selection test. Then, in terms of the uncorrected LR statistic, one would not reject H_0 whenever $-c + n^{-1/2}K_n(F_\theta,G_\gamma)/\hat{\omega}_n \le n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n \le c + n^{-1/2}K_n(F_\theta,G_\gamma)/\hat{\omega}_n$ where c is obtained from the standard normal distribution. It is clear that the main effect of the correction factor $K_n(F_\theta,G_\gamma)$ is to translate the critical region $(-c,+c)$ in the appropriate direction. Which correction factor is preferable depends on how well the exact small sample distribution of $n^{-1/2}\tilde{LR}_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n$ is approximated under H_0 by the asymptotic N(0,1) distribution.

A second fundamental difference between our approach and the previous literature on model selection is that our approach is probabilistic. Though Amemiya (1980) and McAleer and Bera (1983) have argued that an important difference between non-nested hypothesis testing and model selection is that the former framework allows "a probabilistic statement to be made regarding model selection," while the second does not, this criticism no longer applies to our approach which puts model selection in a significance testing situation. Indeed, by appropriately normalizing the LR statistic, we were able to construct a directional test of the hypothesis that the competing models are equivalent against the hypothesis that one of the two models is "better." As a consequence we do not necessarily have to choose a "best" model if the competing models turn out to be statistically equivalent.

Our definitions have the desirable property that a correctly specified model is necessarily at least as good as any other models.

They are nonetheless arbitrary. Indeed, there exist many criteria other than the KLIC that can be used to measure the distance between two distributions. Clearly, an analysis analogous to the one given here can be worked out for each of these other criteria. For instance, using the mean square error (MSE) of prediction, White and Olson (1979) obtained a symmetric and directional normal test for choosing between two non-linear regression models. When the errors are normally distributed, the KLIC and the MSE of prediction lead, however, to identical definitions of equivalence. Moreover, as Lien and Vuong (1986) showed, the White and Olson test and our LR-based test become asymptotically equivalent when the competing models are normal linear regressions.

Finally, one may not be so much interested in the truth of a model, but may be concerned by the number of parameters in a model. To take into account the parsimonious nature of a model, one may add to the criterion (5.1) a penalty $k(\cdot)$ depending on the number of parameters in the model. In this case, the model F_θ is said to be better than, equivalent to, or worse than the competing model G_γ if and only if

$$\Delta \equiv E^0\left[\log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] - [k(p) - k(q)] \tag{5.14}$$

is positive, equal to zero, or negative respectively.[10] Let $\tilde{H}_0$, $\tilde{H}_f$, and $\tilde{H}_g$ denote the hypotheses $\Delta = 0$, $\Delta > 0$, and $\Delta < 0$ respectively. As before we can consider the statistic (5.11) where the correction factor is now:

$$K_n(F_\theta,G_\gamma) = nk(p) - nk(q). \tag{5.15}$$

<u>Theorem 5.4</u> (Alternative Model Selection Tests for Strictly Non-Nested Models): Let $K_n(F_\theta,G_\gamma)$ be as in (5.15). Given Assumptions A1-A6,

(i) under $\tilde{H}_0$: $n^{-1/2}\widetilde{LR}_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n \xrightarrow{D} N(0,1)$,

(ii) under $\tilde{H}_f$: $n^{-1/2}\widetilde{LR}_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n \xrightarrow{a.s} +\infty$,

(iii) under $\tilde{H}_g$: $n^{-1/2}\widetilde{LR}_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n \xrightarrow{a.s.} -\infty$.

Theorem 5.4 generalizes Theorem 5.2 to allow for any kind of penalty function in the definition of equivalent models. As in corollary 5.3, $\tilde{\omega}_n$ can replace $\hat{\omega}_n$ in that theorem. A fundamental difference is that the null and alternative hypotheses are now different from those considered up to now. Also, unlike Corollary 5.3, the correction factor (5.15) does not have to satisfy Condition (5.12). The remarks following Corollary 5.3 nonetheless apply, and for instance, one cannot reject $\tilde{H}_0$ whenever $-c + n^{-1/2}K_n(F_\theta,G_\gamma)/\hat{\omega}_n \leq n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n \leq c + n^{-1/2}K_n(F_\theta,G_\gamma)/\hat{\omega}_n$.

In the next sections on overlapping models and nested models, we shall not discuss the generalizations of Corollary 5.3 and Theorem 5.4. It is clear that similar results can be established.

6. OVERLAPPING MODELS

In this section, we shall apply our model selection approach to the case where the two competing models are overlapping. A simple example of two overlapping models is that of two standard linear regression models with some common explanatory variables. Another example is the dichotomous logit and probit models.[11] As in the previous section, we shall propose some significance tests for discriminating and choosing between two models. We first give a formal definition of overlapping models.

Definition 6.1 (Overlapping Models): Two conditional models F_θ and G_γ are overlapping if and only if:

$$\text{(i)} \quad F_\theta \cap G_\gamma \neq \emptyset, \tag{6.1}$$

$$\text{(ii)} \quad F_\theta \not\subset G_\gamma \text{ and } G_\gamma \not\subset F_\theta. \tag{6.2}$$

Condition (i) says that F_θ and G_γ must have some common conditional distributions for Y_t given Z_t, while condition (ii) says that neither model must be nested in the other.

As in the previous section, our objective is to construct tests of H_0 against H_f or H_g. Given the definitions (5.2)-(5.4) of these hypotheses, a natural test statistic is again the LR statistic. The overlapping case is, however, more difficult than the strictly non-nested case for the following reason. Contrary to the strictly non-nested case, the asymptotic distribution of the LR statistic and the speed at which it converges to the distribution is unknown under

the null hypothesis H_0. Indeed, since $F_\theta \cap G_\gamma \neq \emptyset$, then one may have $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$. From Theorem 3.5, it follows that, under H_0: $E^0[\log f(Y_t|Z_t;\theta_*)] = E^0[\log g(Y_t|Z_t;\gamma^*)]$:

(i) if $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$,

$$2LR_n(\hat\theta_n,\hat\gamma_n) \xrightarrow{D} M_{p+q}(\cdot,\lambda_*), \tag{6.3}$$

(ii) if $f(\cdot|\cdot;\theta_*) \neq g(\cdot|\cdot;\gamma_*)$,

$$n^{-1/2}LR_n(\hat\theta_n,\hat\gamma_n) \xrightarrow{D} N(0,\omega_*^2). \tag{6.4}$$

Since one does not know a priori if $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ holds, one does not know the form of the asymptotic distribution of the LR statistic under the null hypothesis H_0. We distinguish two cases: the general case and the case where one knows a priori that at least one model is correctly specified.

For the general case we propose a sequential procedure which consists in testing first whether $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ and then in using the appropriate null distribution of the LR statistic to construct a model selection test. From Lemma 4.1, we know that $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ if and only if $\omega_*^2 = 0$. Thus, for the first step, a natural test can be based on the variance statistics $\hat\omega_n^2$ and $\tilde\omega_n^2$ of which the asymptotic properties are derived in Section 3. We call such a test, the variance test since it is used to test $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ against $f(\cdot|\cdot;\theta_*) \neq g(\cdot|\cdot;\gamma_*)$, or equivalently:

$$H_0^\omega:\ \omega_*^2 = 0 \quad \text{against} \quad H_A^\omega:\ \omega_*^2 \neq 0.^{12} \tag{6.5}$$

Once it is known whether or not $\omega_*^2 = 0$, then one can use the appropriate null distribution of the LR statistic to test H_0 against H_f or H_g. The second step simplifies since one need not in fact carry out a test of H_0 against H_f or H_g when $\omega_*^2 = 0$. Indeed H_0^ω is included in H_0 since if $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ then the models F_θ and G_γ must necessarily be equivalent. On the other hand, when $\omega_*^2 \neq 0$, then one may have $E^0[\log f(Y_t|Z_t;\theta_*)] = E^0[\log g(Y_t|Z_t;\gamma_*)]$ so that a test of H_0 against H_f or H_g must still be carried out. However, when $\omega_*^2 \neq 0$, then (6.4) holds so that the simple normal test based on $n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n$ or $n^{-1/2}LR_n(\hat{\theta},\hat{\gamma}_n)/\tilde{\omega}_n$ discussed in the previous section can be applied.

To summarize, the sequential procedure is:

(i) Test H_0^ω against H_A^ω using the variance test based on $n\hat{\omega}_n^2$ or $n\tilde{\omega}_n^2$. If H_0^ω cannot be rejected, then conclude that the models F_θ and G_γ cannot be discriminated given the data. If H_0^ω is rejected, then proceed to

(ii) Test H_0 against H_f or H_g using the normal model selection test based on the statistic $n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n$ or $n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\tilde{\omega}_n$ as discussed in Section 5.

As a test of the null hypothesis of interest H_0 that the models are equivalent, this sequential procedure has an exact significance level which is asymptotically bounded above by the maximum of the asymptotic significance levels α_1 and α_2 used for the variance test (i) and the normal LR-test (ii).[13] For instance if $\alpha_1 = \alpha_2 = 10\%$, than the exact significance level of this procedure, as a test of H_0, is asymptotically no larger than 10%.

We now consider in more detail the variance test to be used in the first step. Let $\hat{\lambda}_n$ be the vector of p + q eigenvalues of $\hat{W}_n$ where $\hat{W}_n$ is the sample analog of W as defined in Equation (3.9). For instance, $\hat{W}_n$ is obtained by replacing in Equation (3.9) the matrix $B_{fg}(\theta_*,\gamma_*)$, say, by its sample analog $B_{fgn}(\hat{\theta}_n,\hat{\gamma}_n)$ defined in Equation (2.9). Let $\hat{\lambda}_n^2$ be the vector of squares of $\hat{\lambda}_n$.

<u>Theorem 6.2</u> (Variance Tests for Discrimination): Given Assumptions A1-A7,

(i) under H_0^ω, for any $x \geq 0$,

$$\Pr(n\hat{\omega}_n^2 \leq x) - M_{p+q}(x;\hat{\lambda}_n^2) \overset{a.s.}{\to} 0, \tag{6.6}$$

(ii) under H_A^ω, $n\hat{\omega}_n^2 \overset{a.s.}{\to} +\infty$,

(iii) properties (i) and (ii) hold for $n\tilde{\omega}_n^2$.

The variance test consists first in choosing a critical value x so that $M_{p+q}(x,\hat{\lambda}_n^2) = 1 - \alpha\%$ for some significance level α, and then in rejecting H_0^ω if $n\hat{\omega}_n^2 > x$.[14] Part (i) ensures that the asymptotic size is α, while Part (ii) says that the test is consistent. Similar conclusion applies to the test based on $n\tilde{\omega}_n^2$. Let us note that computation of the statistic $n\hat{\omega}_n^2$ and $n\tilde{\omega}_n^2$ is straightforward given their definitions (4.2) and (4.3).

As mentioned in Section 4, computation of the eigenvalues $\hat{\lambda}_n$ somewhat simplifies if the information matrix equivalences (3.11) hold. Moreover, the eigenvalues $\hat{\lambda}_n$ need not be computed when condition (iii) or (iv) of Theorem 4.4 holds, in which case both $n\hat{\omega}_n^2$ and $n\tilde{\omega}_n^2$ converges, under H_ω^0, to a chi-square distribution with degrees of freedom equal to $p + q - 2$ rank $B_{fgn}(\hat{\theta}_n,\hat{\gamma}_n)$. As mentioned in Section 4, condition (iii) - (iv) of Theorem 4.4 are satisfied when F_θ and G_γ are orthogonal models, in which case both $n\hat{\omega}_n^2$ and $n\tilde{\omega}_n^2$ converge to a chi-square distribution with $p + q$ degrees of freedom under the null hypothesis H_0^ω.

As pointed out earlier, the difficulty in selecting among overlapping models arises from the fact that $f(\cdot|\cdot;\theta_*)$ may or may not be equal to $g(\cdot|\cdot;\gamma_*)$ under the null hypothesis H_0: $E^0[\log f(Y_t|Z_t;\theta_*)] = E^0[\log g(Y_t|Z_t;\gamma_*)]$ so that the form of the asymptotic null distribution of the LR statistic is a priori unknown. This is not, however, the case if one knows a priori that at least one of the two overlapping models is correctly specified, as this is frequently assumed in the model selection literature. Let us note that we do not say whether it is F_θ or G_γ that is correctly specified.

<u>Lemma 6.3</u>: Given Assumptions A2 and A3, suppose that

$$H^0(y|z) \in F_\theta \cup G_\gamma, \tag{6.7}$$

then the following statements are equivalent:

(i) $H^0(y|z) \in F_\theta \cap G_\gamma$,

(ii) $f(\cdot|\cdot,\theta_*) = g(\cdot|\cdot;\gamma_*)$,

(iii) $E^0[\log f(Y_t|Z_t;\theta_*)] = E^0[\log g(Y_t|Z_t;\gamma_*)]$.

From (i) and (iii) it follows that, when at least one model is known to be correctly specified, then the models F_θ and G_γ are (KLIC) equivalent if and only if the other model is correctly specified. That (i) implies (iii) is obvious. The intuition behind the reverse implication is based on the fact that when the model F_θ, say, is correctly specified then $E^0[\log f(Y_t|Z_t;\theta_*)] = E^0[\log h^0(Y_t|Z_t)]$. Thus, when condition (iii) holds, $E^0[\log g(Y_t|Z_t;\gamma_*)] = E^0[\log h^0(Y_t|Z_t)]$ and therefore G_γ must be correctly specified.

From (ii) and (iii) we have that the models F_θ and G_γ are equivalent if and only if $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$.[15] The importance of this second equivalence is that under the null hypothesis H_0, we now always have $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ so that the asymptotic distribution of the LR statistic is given by the weighted sum of chi-squares obtained in Theorem 3.5 - (i). Thus in this case we can bypass the above sequential procedure, and directly construct a model selection test based on the LR statistic.

<u>Theorem 6.4</u> (Model Selection Test for Overlapping Models): In addition to Assumptions A1-A5, suppose that at least one model is correctly specified.[16] Then:

(i) under H_0, for any $x \geq 0$,

$$\Pr(2LR_n(\hat{\theta}_n,\hat{\gamma}_n) \leq x) - M_{p+q}(x;\hat{\lambda}_n) \xrightarrow{a.s.} 0, \tag{6.8}$$

(ii) under H_f: $2LR_n(\hat{\theta}_n,\hat{\gamma}_n) \xrightarrow{a.s.} +\infty$,

(iii) under H_g: $2LR_n(\hat{\theta}_n,\hat{\gamma}_n) \xrightarrow{a.s.} -\infty$.

The LR-based test is carried out by choosing critical values from the weighted sum of chi-squares $M_{p+q}(\cdot;\hat{\lambda}_n)$. Since the LR-based test is two sided, two critical values c_1 and c_2 are chosen, one from the upper-tail and one from the lower-tail of this distibution. As for the normal LR-based test of Section 5, the test is directional in the sense that H_0 is rejected in favor of H_f or H_g according to whether $2LR_n(\hat{\theta}_n,\hat{\gamma}_n) > c_1$ or $2LR_n(\hat{\theta}_n,\hat{\gamma}_n) < c_2$ respectively.[17]

Let us also note that the burdensome computation of the eigenvalues $\hat{\lambda}_n$ simplifies when one model is correctly specified. Indeed, under H_0, when one model is correctly specified then the other must also be correctly specified (see Lemma 6.3) so that, from the information matrix equivalences (3.11), the matrix W reduces to:[18]

$$W = \begin{bmatrix} I_p & B_{fg}(\theta_*,\gamma_*)B_g^{-1}(\gamma_*) \\ -B_{gf}(\gamma_*,\theta_*)B_f^{-1}(\theta_*) & -I_q \end{bmatrix}. \tag{6.9}$$

In addition, the eigenvalues $\hat{\lambda}_n$ need not be computed when the two overlapping models are orthogonal in which case the off-diagonal blocks of W are identically null. The distribution then reduces to the distribution of a difference between two independent chi-squares with p and q degrees of freedom.

7. NESTED MODELS

We now consider the more familiar case of nested models. We first relate our probabilistic model selection approach to the classical nested-hypothesis testing situation. Then we propose a LR-based test for selecting between two nested models. This test reduces to the classical Neyman-Pearson (1928) LR test when the largest model is correctly specified. We also propose a new test for nested hypotheses based on the variance statistics of Section 3.

We first give a formal definition of nested models.

<u>Definition 7.1</u> (Nested Models): Two conditional models F_θ and G_γ are nested if and only if:

$$G_\gamma \subset F_\theta \text{ or } F_\theta \subset G_\gamma. \tag{7.1}$$

We shall assume throughout this section that G_γ is nested in F_θ, i.e., that $G_\gamma \subset F_\theta$. We make the following regularity assumption on the parameterizations θ and γ.

<u>Assumption A8</u>: There exists a C^2-function $\phi(\cdot)$ from Γ to Θ such that:

$$g(\cdot|\cdot;\gamma) = f(\cdot|\cdot;\phi(\gamma)) \text{ for any } \gamma \text{ in } \Gamma. \tag{7.2}$$

Condition (7.2) states that any conditional density $g(\cdot|\cdot;\gamma)$ is also a conditional density $f(\cdot|\cdot;\theta)$ for some θ in Θ. Since $\phi(\Gamma)$ is included in Θ, then the conditional model G_γ is indeed nested in F_θ.

Let us note that the pseudo-true parameter θ_* is not necessarily equal to $\phi(\gamma_*)$ since θ_* may not even belong to $\phi(\Gamma)$. The next result relates the condition $\theta_* \in \phi(\Gamma)$ to the condition that F_θ and G_γ are equivalent, and to the condition that $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$.

<u>Lemma 7.2</u>: Given Assumptions A2, A3, and A8, the following statements are equivalent:

(i) $\theta_* = \phi(\gamma_*)$,

(ii) $\theta_* \in \phi(\Gamma)$,

(iii) $E^0[\log f(Y_t|Z_t;\theta_*)] = E^0[\log g(Y_t|Z_t;\gamma_*)]$,

(iv) $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$.

Lemma 7.2 is important since it shows that our model selection approach coincides with the classical testing approach when the models are nested. For, the condition H_0^θ: $\theta_* \in \phi(\Gamma)$ can be interpreted as the condition that θ_* satisfies some restrictions,and thus corresponds to the parametric null hypothesis of the classical testing framework in <u>implicit</u> form. On the other hand, the null hypothesis in our model selection approach is H_0. From (ii) and (iii), we have that H_0^θ and H_0 are equivalent, as must be their respective alternatives $H_A^\theta : \theta_* \notin \phi(\Gamma)$ and $H_f \cup H_g$. Thus testing H_0^θ against H_A^θ is equivalent to testing H_0 against $H_f \cup H_g$. In other words, testing whether or not θ_* satisfies some restrictions is equivalent to testing whether or not the smaller model is equivalent to the larger model.[19]

As a matter of fact, the alternative to the null hypothesis H_0 is H_f, i.e., that the model F_θ is better than G_γ. Indeed F_θ can never be worse that G_γ since we must have:

$$E^0[\log f(Y_t|Z_t;\theta_*)] \geq E^0[\log g(Y_t|Z_t;\gamma_*)], \tag{7.3}$$

so that H_g can never occur. Thus, we in fact have the equivalence

between H_A^θ and H_f.

As argued earlier, the LR statistic is a natural statistic for selecting among models. Thus, we shall consider a LR-based test of H_0 against H_f or equivalently of H_0^θ against H_A^θ. From Lemma 7.2, we always have $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ under the null hypothesis H_0. Thus, there is here no ambiguity as to the asymptotic distribution of the LR statistic which is the weighted sum of chi-squares obtained in Theorem 3.5 - (i). We need a preliminary result relating the matrices B_g, A_g, B_f, A_f and B_{fg} under the null hypothesis H_0.

<u>Lemma 7.3</u>: Given Assumptions A2 - A5, and A8, then under H_0^θ:

(i) $B_g(\gamma_*) = \dfrac{\partial\phi'(\gamma_*)}{\partial\gamma} B_f(\theta_*) \dfrac{\partial\phi(\gamma_*)}{\partial\gamma'}$; $A_g(\gamma_*) = \dfrac{\partial\phi'(\gamma_*)}{\partial\gamma} A_f(\theta_*) \dfrac{\partial\phi(\gamma_*)}{\partial\gamma'}$,

(ii) $B_{gf}(\gamma_*,\theta_*) = \dfrac{\partial\phi'(\gamma_*)}{\partial\gamma} B_f(\theta_*)$,

(iii) $q \leq p$, rank $\dfrac{\partial\phi'(\gamma_*)}{\partial\gamma} = q$.

Let us note that Lemma 7.3 says in particular that the dimension q of the parameters γ cannot be greater than the dimension p of the parameter θ. This is expected since G_γ is nested in F_θ.

It is convenient to define $\tilde\theta_n \equiv \phi(\hat\gamma_n)$; $\tilde\theta_n$ is nothing else than the constrained (quasi) maximum likelihood estimator of θ_* subject to the constraints that θ belongs to $\phi(\Gamma)$. Then the usual LR statistic of the unconstrained vs. the constrained model is:

$$LR_n(\hat\theta_n,\tilde\theta_n) \equiv LR_n(\hat\theta_n,\hat\gamma_n),$$

$$= \sum_{t=1}^{n} \log \frac{f(Y_t|Z_t;\hat{\theta}_n)}{f(Y_t|Z_t;\tilde{\theta}_n)}, \tag{7.4}$$

where the second equality follows from Assumption A8 and the definition of $\tilde{\theta}_n$.

The next result is similar to Kent (1982) Theorem 3.1, and gives the properties of the model selection or nested hypothesis test based on the LR statistic. In particular, it greatly simplifies the computation of the non-zero eigenvalues of the general matrix W in Theorem 3.5 by replacing W by a matrix $\underline{W}$ of lower dimension. Specifically, let:

$$\underline{W} = B_f(\theta_*)\left[\frac{\partial \phi(\gamma_*)}{\partial \gamma'} A_g^{-1}(\gamma_*)\frac{\partial \phi'(\gamma_*)}{\partial \gamma} - A_f^{-1}(\theta_*)\right], \tag{7.5}$$

and let $\hat{\lambda}_n$ be the vector of p eigenvalues of the sample analog $\underline{W}_n$ of $\underline{W}$.

<u>Theorem</u> <u>7.4</u> (LR Tests for Nested Models): Given Assumptions A1-A5 and A8, the eigenvalues $\hat{\lambda}_n$ are almost surely all real non-negative and:

(i) under H_0^θ, for any $x \geq 0$,

$$\Pr(2LR_n(\hat{\theta}_n,\tilde{\theta}_n) \leq x) - M_p(x;\hat{\lambda}_n) \overset{a.s.}{\to} 0, \tag{7.6}$$

(ii) under H_A^θ, $2LR_n(\hat{\theta}_n,\tilde{\theta}_n) \overset{a.s.}{\to} +\infty$.

The test is one sided. It is carried out by choosing a critical value from $M_p(\cdot;\hat{\lambda}_n)$ and by rejecting the hypothesis that the models are equivalent or that θ^* belongs to $\phi(\Gamma)$ if twice the LR statistic is greater than this critical value. The test applies whether or not the larger model is correctly specified.

As noted by White (1982a), if the information matrix holds for the larger model then one obtains from Lemma 7.3 and Theorem 3.6:

<u>Corollary</u> <u>7.5</u> (LR Tests for Nested Models given Information Matrix Equivalence): Given Assumptions A1-A5, A8 suppose that $A_f(\theta_*) + B_f(\theta_*) = 0$:

(i) under H_0^θ, $2LR_n(\hat{\theta}_n,\tilde{\theta}_n) \overset{D}{\to} \chi^2_{p-q}$,

(ii) under H_A^θ, $2LR_n(\hat{\theta}_n,\tilde{\theta}_n) \overset{a.s.}{\to} +\infty$.

The well-known Wilks (1938) result follows since the information matrix equivalence $A_f(\theta_*) + B_f(\theta_*) = 0$ holds if the larger model is correctly specified (see footnote 18).

Using the equivalence between H_0^θ and H_0, we have motivated the LR statistic as a basis for constructing a test of H_0^θ against H_A^θ under general conditions. But from Lemmas 7.2 and 4.1, we also have the equivalence between H_0^θ and $H_0^\omega : \omega_*^2 = 0$. This suggests that, to test the parametric hypothesis H_0^θ against H_A^θ we can equivalently test H_0^ω against H_A^ω.

Thus, we have a new test for nested hypothesis based on the variance statistics $\hat{\omega}_n^2$ and $\tilde{\omega}_n^2$ as defined in Equations (4.2) and (4.3). Let $\hat{\lambda}_n^2$ be the squares of the eigenvalues $\hat{\lambda}_n$.

__Theorem 7.6__ (Variance Tests for Nested Models): Given Assumptions A1–A8:

(i) under H_0^θ, for any $x \geq 0$,

$$\Pr(n\hat{\omega}_n^2 \leq x) - M_p(x;\hat{\lambda}_n^2) \xrightarrow{a.s.} 0, \tag{7.7}$$

(ii) under H_A^θ, $n\hat{\omega}_n^2 \xrightarrow{a.s.} +\infty$,

(iii) properties (i) and (ii) hold for $n\tilde{\omega}_n^2$.

As for the LR test of Theorem 7.4, variance tests are one-sided. They are carried out by choosing a critical value from $M_p(\cdot;\hat{\lambda}_n^2)$ and by rejecting the hypothesis that θ_* belongs to $\phi(\Gamma)$ if $n\hat{\omega}_n^2$ or $n\tilde{\omega}_n^2$ is larger than this critical value. These statistics $n\hat{\omega}_n^2$ and $n\tilde{\omega}_n^2$ are readily computed. Indeed from Equation (4.2) and (4.3) we have using Assumption A8:

$$n\hat{\omega}_n^2 = \sum_{t=1}^{n}\left[\log\frac{f(Y_t|Z_t;\hat{\theta}_n)}{f(Y_t|Z_t;\tilde{\theta}_n)}\right]^2 - \frac{1}{n}LR_n^2(\hat{\theta}_n,\tilde{\theta}_n), \tag{7.8}$$

$$n\tilde{\omega}_n^2 = \sum_{t=1}^{n}\left[\log\frac{f(Y_t|Z_t;\hat{\theta}_n)}{f(Y_t|Z_t;\tilde{\theta}_n)}\right]^2, \tag{7.9}$$

where $\tilde{\theta}_n$ is the constrained ML estimator. For instance, $n\hat{\omega}_n^2$ is the sum of square residuals in a linear regression of $m_t \equiv \log[f(Y_t|Z_t;\hat{\theta}_n)/f(Y_t|Z_t;\tilde{\theta}_n)]$ on the constant term.[20]

If, however, the larger model is correctly specified, then the

eigenvalues $\hat{\lambda}_n$ need not be computed since in this case the limiting distribution reduces to the central chi-square distribution with $p - q$ degrees of freedom, as other classical statistics.

__Corollary 7.7__ (Variance Tests for Nested Models given Information Matrix Equivalence): Given Assumptions A1 – A8, suppose that $A_f(\theta_*) + B_f(\theta_*) = 0$:

(i) under H_0^θ, $n\hat{\omega}_n^2 \xrightarrow{D} \chi^2_{p-q}$,

(ii) under H_A^θ, $n\hat{\omega}_n^2 \xrightarrow{a.s.} +\infty$,

(iii) properties (i) and (ii) hold for $n\tilde{\omega}_n^2$.

As mentioned earlier, the information matrix equivalence $A_f(\theta_*) + B_f(\theta_*) = 0$ holds if the larger model is correctly specified.

8. CONCLUSION

In this paper, we have proposed a new and general approach to model selection whether the competing models are nested, overlapping or non-nested, and whether the models are correctly specified. This approach has the desirable property that it coincides with the usual classical testing approach when the models are nested. It is probabilistic and is based on testing if the competing models are as close to the true distribution against the hypothesis that one model is closer than the other. Since the maximum log-likelihood of a model is a natural estimator of the distance between the model and the true distribution as measured by the Kullback-Leibler information criterion, all our model selection tests, with the exception of the variance tests discussed above, are LR-based tests. As a prerequisite, we have therefore fully characterized the asymptotic distribution of the LR statistic under the most general conditions.

In Section 5 on non-nested models, we have contrasted our model selection approach to the more familiar one originated by Akaike (1973, 1974). In Section 7 on nested models, we have shown that classical nested hypothesis tests are in fact model selection tests. We now express our view on the general purpose of model selection, specification testing, and non-nested hypothesis testing in econometric modelling.

First, it is important to note that model selection tests, as we have defined, can be thought of as specification tests. Indeed, given a statistical model, it is natural to question its validity. If one has in mind some reasons for possible misspecification of the initial model, one has in fact a list of competing models. To simplify, suppose that there is only one competing model. Then, by carrying out the model selection tests proposed in this paper, one may be able to conclude that the initial model is misspecified. Specifically, if one rejects the equivalence between these two models in favor of the competing model being better, then the initial model must be misspecified. Moreover, rejection suggests in which direction the initial model must be modified since the test indicates that the alternative model is closer to the truth.[21] On the other hand, in the other two situations where the equivalence cannot be rejected or the equivalence is rejected in favor of the initial model being better, one cannot infer that the initial model is correctly specified. This is usual in specification testing where acceptance of the null hypothesis does not in general imply correct specification of the model under test.

The previous paragragh does not imply that specification tests as originated by Hausman (1978) and White (1982a) are unimportant.[22] First, as we have seen in the overlapping case, our model selection tests simplify if the information matrix equivalence holds or if at least one model is correctly specified. Second, and more importantly, specification tests are useful when one does not have any precise alternative models in mind. There is, however, a difference between model specification testing and our approach to model selection. Indeed, in model specification testing, one first performs various available specification tests, and then investigates the power of the tests so as to interpret the implicit alternatives to the initial model specification. On the other hand, in model selection, one must first have some ideas about possible form of misspecification to formulate alternative models. Then one carries out some model selection tests to decide if the initial model is correctly specified.[23]

We now turn to the important comparison between our model selection approach and the non-nested hypotheses approach as originated by Cox (1961, 1962). In the conditional framework of Section 2, the Cox statistic for testing the model F_θ using the evidence provided by G_γ is based on the <u>modified</u> LR statistic:

$$T_n^f = \frac{1}{n}LR_n(\hat{\theta}_n,\hat{\gamma}_n) - \frac{1}{n}\sum_{t=1}^{n}\int_Y \log \frac{f(y|Z_t;\hat{\theta}_n)}{g(y|Z_t;\hat{\gamma}_n)} f(y|Z_t;\hat{\theta}_n)dy.^{24} \qquad (8.1)$$

It is easy to see that the implicit null and alternative hypotheses of the Cox test are:

$$H_0^f : \int_Z \{\int_Y \log \frac{f(y|z;\theta_*)}{g(y|z;\gamma_*)}[h^0(y|z) - f(y|z;\theta_*)]dy\}h^0(z)dz = 0, \qquad (8.2)$$

where $h^0(z)$ is the true marginal density of Z_t, and H_A^f is the negation of H_0^f. It is clear that if F_θ is correctly specified so that $h^0(y|z) = f(y|z;\theta_*)$, then Equation (8.2) is satisfied. On the other hand, the null hypothesis H_0^f may hold even though the model F_θ is misspecified so that the Cox-test does not have power against this type of misspecification. Along the same lines, let us note that when G_γ is nested in F_θ, the parametric hypothesis H_0^θ: $\theta_* = \phi(\gamma_*)$ is included but not necessarily equal to H_0^f. Hence, contrary to our approach, Cox's approach does not coincide with the classical hypothesis approach when the models are nested. This is so because Cox's null hypothesis H_0^f is different from our null hypothesis H_0.

Though MacKinnon (1983) has argued that non-nested hypothesis tests should be interpreted as "model specification tests using the evidence provided by non-nested alternative hypotheses," it is well-known that Cox-type tests have also been used as discrimination or model selection tests. This is done by reversing the role of F_θ and G_γ in which case one has nine possible outcomes (see, e.g., Fisher and McAleer (1979)) according to whether H_-^f, H_0^f, or H_+^f holds on the one hand, and H_-^g, H_0^g, or H_+^g holds on the other hand. The hypotheses H_-^f, H_0^f, and H_+^f corresponds to whether the left-hand side of (8.2) is negative, zero, or positive. Similar definitions apply to H_-^g, H_0^g, and H_+^g when F_θ is replaced by G_γ. Given our definitions of equivalent and better models, we can provide in the following table the conclusion associated with each of these nine possibilities:

	H_-^f	H_0^f	H_+^f
H_-^g	indecisive	$F_\theta \geq G_\gamma$	$F_\theta > G_\gamma$
H_0^g	$G_\gamma \geq F_\theta$	$f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ $(\Rightarrow F_\theta \equiv G_\gamma)$	impossible
H_+^g	$G_\gamma > F_\theta$	impossible	impossible

where, for instance, $G_\gamma \geq F_\theta$ indicates that G_γ is at least as good as F_θ, and $G_\gamma > F_\theta$ indicates that G_γ is (strictly) better than F_θ.

We now explain such a table which relies on the remark that the hypotheses H_-^f, H_0^f, and H_+^f can be rewritten respectively as:

$$E^0[\log f(Y_t|Z_t;\theta_*)] \lesseqgtr E^0[\log g(Y_t|Z_t;\gamma_*)] + \int_Z \int_Y \log \frac{f(y|z;\theta_*)}{g(y|z;\gamma_*)} f(y|z;\theta_*)h^0(z)dydz. \qquad (8.3)$$

Similarly, H_-^g, H_0^g, and H_+^g can be rewritten as:

$$E^0[\log g(Y_t|Z_t;\gamma_*)] \lesseqgtr E^0[\log f(Y_t|Z_t;\gamma_*)]$$

$$+ \int_Z \int_Y \log \frac{g(y|z;\gamma_*)}{f(y|z;\theta_*)} g(y|z;\gamma_*) h^0(z) dy dz. \qquad (8.4)$$

By Jensen's inequality, the second terms in Equations (8.3) and (8.4) are both non-negative, and equal to zero if and only if $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$. This explains why the three possibilities (H_0^g, H_+^f), (H_+^g, H_0^f), and (H_+^g, H_+^f) cannot occur (asymptotically).[25] As a consequence, when one rejects, say, H_0^f in favor of H_+^f in a Cox test, one need not reverse the hypotheses since one already knows that F_θ is better than G_γ. Moreover, from the second column of the table, one need not either reverse the hypotheses when H_0^f cannot be rejected since F_θ is at least as good as G_γ. This follows by noticing from Equations (8.3) and (8.4) that (i) (H_-^g, H_0^f) implies that F_θ is at least as good as G_γ, (ii) (H_0^g, H_0^f) implies that $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ and hence that F_θ and G_γ are equivalent, and (iii) that (H_+^g, H_0^f) cannot occur. But let us note that if one cannot reject H_0^f so that $F_\theta \geq G_\gamma$, there is no way using the Cox test to determine if F_θ is (strictly) better than G_γ. The situation becomes worse if H_0^f is rejected in favor of H_-^f. Indeed, as the first column of the table indicates, even if one reverse the hypotheses, one may conclude that the combination (H_-^g, H_-^f) holds, but this combination is indecisive since all we know is that $E^0[\log f(Y_t|Z_t;\theta_*)] - E^0[\log g(Y_t|Z_t;\gamma_*)]$ is less than the second term in Equation (8.3), but larger than minus the second term in Equation (8.4).

Though non-nested hypothesis tests have sometimes been advocated by the fact than "an economic researcher would be more

interested in the truth of a particular model than in choosing from among a given set of models" (Datsoor (1981)), we believe that this leads to a non-optimal strategy in econometric modeling.[26] Indeed, instead of testing the specification of each model in a list of competing models using the evidence provided by the alternative models, as this is done in non-nested hypothesis testing, it is more economical to choose the best model among this list and then, if one is still interested in the truth, to perform either some specification tests on the best model or to expand the list of competing models so as to perform some further model selection tests. That this latter strategy is internally consistent is ensured by the fact that our definition of a "best" model is compatible with that of a model being correctly specified.

Much work remains to be done. First, an important task is to apply the proposed tests for model selection to some special cases such as the linear and non-linear regression models. Comparison between the resulting tests and the available Cox-type tests would be useful. Second, asymptotic power comparison between our model selection tests, when treated as model specification tests, and current specification tests would be interesting. Third, it would be useful to compare our approach to the comprehensive approach advocated by Atkinson (1969, 1970) which requires to nest the competing models in a larger model. An interesting case is that of a linear as a log-linear functional form as considered by Box and Cox (1964). Fourth, it would be interesting to compare the performance of our model selection tests to the tests using the encompassing principle as advocated by Hendry (1983), and Mizon and Richard (1982). Fifth, the above model selection tests have been obtained under the assumption that there are only two competing models. It is therefore important to generalize our procedures to the case where there are many competing models. It appears that the likelihood ratio principle can still be invoked by taking the supremum of the log-likelihood over all the alternative models.

APPENDIX

Except when explicitly mentioned, all the matrices A_f, B_f, A_g, B_g, and B_{fg} are evaluated at the pseudo-true values θ_* and γ_*.

Proof of Lemma 2.1: Given Assumptions A1-A5, we obtain using the Taylor expansions of the normal equations:

$$0 = n^{-1/2}\frac{\partial L_n^f(\theta_*)}{\partial\theta} + A_f \cdot n^{1/2}(\hat{\theta}_n - \theta_*) + o_p(1), \tag{A.1}$$

$$0 = n^{-1/2}\frac{\partial L_n^g(\gamma_*)}{\partial\gamma} + A_g \cdot n^{1/2}(\hat{\gamma}_n - \gamma_*) + o_p(1), \tag{A.2}$$

(see, e.g., Vuong (1983), proof of Theorem 3)). On the other hand from the multivariate Central Limit Theorem (see, e.g. Rao (1973)):

$$n^{-1/2}\begin{bmatrix} \partial L_n^f(\theta_*)/\partial\theta \\ \partial L_n^g(\gamma_*)/\partial\gamma \end{bmatrix} \xrightarrow{D} N(0, \begin{bmatrix} B_f & ; & B_{fg} \\ B_{gf} & ; & B_g \end{bmatrix}). \tag{A.3}$$

The desired result follows from (A.1) - (A.3) by noticing that A_f and A_g are non-singular (see, White (1982a, Theorem 3.1)).

Proof of Lemma 3.1: Obvious from, e.g., Vuong (1983, Theorem 1).

Proof of Lemma 3.2: Taking a Taylor expansion of $L_n^f(\theta_*)$ around $\hat{\theta}_n$, we have for some $\bar{\theta}_n$ in the segment $[\theta_*, \hat{\theta}_n]$:

$$L_n^f(\theta_*) = L_n^f(\hat{\theta}_n) + \frac{\partial L_n^f(\hat{\theta}_n)}{\partial\theta'}(\theta_* - \hat{\theta}_n) + \frac{1}{2}(\hat{\theta}_n - \theta_*)'\frac{\partial^2 L_n^f(\bar{\theta}_n)}{\partial\theta\partial\theta'}(\hat{\theta}_n - \theta_*). \tag{A.4}$$

But the second term is null by definition of $\hat{\theta}_n$. Since $n^{-1}\partial^2 L_n^f(\bar{\theta}_*)/\partial\theta\partial\theta' = A_f + o_p(1)$, if follows that:

$$L_n^f(\theta_*) = L_n^f(\hat{\theta}_n) + \frac{n}{2}(\hat{\theta}_n - \theta_*)' A_f(\hat{\theta}_n - \theta_*) + o_p(1). \tag{A.5}$$

Similarly, we have:

$$L_n^g(\gamma_*) = L_n^g(\hat{\gamma}_n) + \frac{n}{2}(\hat{\gamma}_n - \gamma_*)' A_g(\hat{\gamma}_n - \gamma_*) + o_p(1). \tag{A.6}$$

Since $LR_n(\theta_*,\gamma_*) = L_n^f(\theta_*) - L_n^g(\gamma_*)$, we obtain:

$$LR_n(\hat{\theta}_n,\hat{\gamma}_n) = LR_n(\theta_*,\gamma_*) - \frac{n}{2}(\hat{\theta}_n - \theta_*)' A_f(\hat{\theta}_n - \theta_*) + \frac{n}{2}(\hat{\gamma}_n - \gamma_*)' A_g(\hat{\gamma}_n - \gamma_*) + o_p(1). \tag{A.7}$$

Part (i) follows from the fact that $LR_n(\theta_*,\gamma_*) = 0$ if $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$. On the other hand, if $f(\cdot|\cdot;\theta_*) \neq g(\cdot|\cdot;\gamma_*)$, then $LR_n(\theta_*,\gamma_*)$ is not zero. But we always have $n^{1/2}(\hat{\theta}_n - \theta_*)$ and $n^{1/2}(\hat{\gamma}_n - \gamma_*)$ being $O_p(1)$. This establishes Part (ii).

Proof of Lemma 3.4: From Moore (1978, Theorem 1), we know that $Y'QY \sim M_m(\cdot;\lambda)$ where λ are the eigenvalues of $\Omega^{1/2}Q\Omega^{1/2}$ where $\Omega^{1/2} = P'D^{1/2}P$, and P is an orthogonal matrix that diagonalizes Ω into D, i.e., $P\Omega P' = D$ and $PP' = P'P = I_m$. It remains to show that the eigenvalues of $\Omega^{1/2}Q\Omega^{1/2}$ are the eigenvalues of $Q\Omega$. Let us order the eigenvalues and eigenvectors so that:

$$D = \begin{bmatrix} D_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad P' = [P_1'; P_0'], \tag{A.8}$$

where D_1 is an $r \times r$ diagonal matrix of which all the diagonal elements are strictly positive (since Ω is p.s.d.). Then, using the orthogonality of P and the properties of determinants, the eigenvalues of $\Omega^{1/2}Q\Omega^{1/2}$ solve:

$$0 = |D^{1/2}PQP'D^{1/2} - \lambda I_m|$$

$$= |D_1^{1/2}P_1QP_1'D_1^{1/2} - \lambda I_r|\lambda^{m-r}. \tag{A.9}$$

Similarly the eigenvalues of $Q\Omega$ solve:

$$0 = |PQP'D - \lambda I_m|$$

$$= |P_1QP_1'D_1 - \lambda I_r|\lambda^{m-r}, \tag{A.10}$$

which is equivalent to (A.9) by pre and post multiplying by $D_1^{1/2}$ and $D_1^{-1/2}$.

Proof of Theorem 3.5: Part (i) follows from Lemma 2.1, Lemma 3.2 - (i) and Lemma 3.4 by considering the quadratic form associated with the block-diagonal matrix:

$$Q = \begin{bmatrix} -A_f & 0 \\ 0 & A_g \end{bmatrix} \tag{A.11}$$

Then, one can check that $Q\sum$ is equal to W as given in Equation (3.9).

From Lemma 3.2 - (ii), we have:

$$n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n) - n^{1/2}E^0\left[\log \frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right]$$

$$= n^{1/2}\left[\frac{1}{n}LR_n(\theta_*,\gamma_*) - E^0\left[\log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right]\right] + o_p(1).$$

But from the multivariate Central Limit Theorem, the first term in the right hand side converges in distribution to $N(0,\omega_*^2)$ where ω_*^2 is the variance defined in Equation (3.6). This variance ω_*^2 is finite given Assumption A6 and the Cauchy-Shwartz inequality. Part (ii) follows.

<u>Proof of Theorem 3.6</u>: From Lemma 3.2 - (i), $2LR_n(\hat{\theta}_n,\hat{\gamma}_n)$ is asymptotically distributed as a quadratic form in $n^{1/2}(\hat{\theta}_n' - \theta_*',\hat{\gamma}_n' - \gamma_*')'$ which is asymptotically normal $N(0,\sum)$ (Lemma 2.1). Thus, from Rao and Mitra (1971, Theorem 9.2.1), $2LR_n(\hat{\theta}_n,\hat{\gamma}_n)$ is asymptotically distributed as a (central) chi-square if and only if:

$$\sum Q\sum Q\sum = \sum Q\sum, \tag{A.12}$$

where Q is given in (A.11), in which case the number of degrees of freedom is tr $Q\sum$.

We now use the fact that:

$$\sum = A^{-1}BA^{-1}, \tag{A.13}$$

where

$$B = \begin{bmatrix} B_f & B_{fg} \\ B_{gf} & B_g \end{bmatrix} ; \quad A = \begin{bmatrix} A_f & 0 \\ 0 & A_g \end{bmatrix}. \tag{A.14}$$

Noticing that $A^{-1}QA^{-1} = Q^{-1}$, Condition (A.14) becomes:

$$BQ^{-1}BQ^{-1}B = BQ^{-1}B. \tag{A.15}$$

Using now the information matrix equivalences (3.11), we obtain after some matrix multiplications that (A.15) is equivalent to:

$$\begin{bmatrix} B_f - B_{fg}B_g^{-1}B_{gf} & ; & B_{fg}B_g^{-1}(B_g - B_{gf}B_f^{-1}B_{fg}) \\ (B_g - B_{gf}B_f^{-1}B_{fg})B_g^{-1}B_{gf} & ; & B_g - B_{gf}B_f^{-1}B_{fg} \end{bmatrix}$$

$$= \begin{bmatrix} B_f - B_{fg}B_g^{-1}B_{gf} & ; & 0 \\ 0 & ; & -B_g + B_{gf}B_f^{-1}B_{fg} \end{bmatrix} \tag{A.16}$$

Hence $2LR_n(\hat{\theta}_n,\hat{\gamma}_n)$ has a limiting (central) chi-square distribution if and only if (3.12) holds. Its number of degrees of freedom is then:

$$\text{tr } Q\sum = \text{tr}\begin{bmatrix} I_p & B_{fg}B_g^{-1} \\ -B_{gf}B_f^{-1} & -I_q \end{bmatrix} = p - q. \tag{A.17}$$

<u>Proof of Lemma 4.1</u>: From Definition (3.7) of ω_*^2, it follows that $\omega_*^2 = 0$ if and only if:

$$\log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)} \overset{a.s.}{=} E^0\left[\log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] = \text{constant},$$

i.e., if and only if: $f(\cdot|\cdot;\theta_*) = Kg(\cdot|\cdot;\gamma_*)$ for some constant K. Since $f(\cdot|\cdot;\theta_*)$ and $g(\cdot|\cdot;\gamma_*)$ are densities, then $K = 1$.[27]

<u>Proof of Lemma 4.2</u>: Given Assumptions A1-A3, and A6, it follows from the Cauchy-Scwartz inequality and Jennrich's uniform Strong Law of Large Numbers (1969, Theorem 2) that

$$\frac{1}{n}\sum_{t=1}^{n}\left[\log \frac{f(Y_t|Z_t;\theta)}{g(Y_t|Z_t;\gamma)}\right]^2 \overset{a.s.}{\to} E^0\left[\log \frac{f(Y_t|Z_t;\theta)}{g(Y_t|Z_t;\gamma)}\right]^2, \tag{A.18}$$

uniformly in θ on Θ. The result follows from Lemma 3.1 and the strong consistency of $\hat{\theta}_n$ and $\hat{\gamma}_n$ to θ_* and γ_*.

<u>Proof of Theorem 4.3</u>: Since $\omega_*^2 = 0$ is equivalent to $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ (Lemma 4.1), it follows from Theorem 3.5 - (i) that: $LR_n(\hat{\theta}_n,\hat{\gamma}_n) = O_p(1)$. Thus, from Equation (4.4), we have:

$$n\tilde{\omega}_n^2 = n\hat{\omega}_n^2 + n^{-1}O_p(1) = n\hat{\omega}_n^2 + o_p(1).$$

Hence, we need only to study the null asymptotic distribution of $n\tilde{\omega}_n^2$.

Using a Taylor expansion around (θ_*,γ_*), we obtain:

$$\tilde{\omega}_n^2 = \frac{1}{n}\sum_{t=1}^{n}\left[\log \frac{f_t(\theta_*)}{g_t(\gamma_*)}\right]^2 + 2\left[\frac{1}{n}\sum_{t=1}^{n}\left[\log \frac{f_t(\theta_*)}{g_t(\gamma_*)}\right]\frac{\partial \log f_t(\theta_*)}{\partial\theta'}\right](\hat{\theta}_n - \theta_*)$$

$$- 2\left[\frac{1}{n}\sum_{t=1}^{n}\left[\log \frac{f_t(\theta_*)}{g_t(\gamma_*)}\right]\frac{\partial \log g_t(\gamma_*)}{\partial\gamma'}\right](\hat{\gamma}_n - \gamma_*)$$

$$+ (\hat{\theta}_n' - \theta_*', \hat{\gamma}_n' - \gamma_*')\bar{V}_n(\hat{\theta}_n' - \theta_n', \hat{\gamma}_n' - \gamma_*')' \tag{A.19}$$

where, to simplify the notation, we have used $f_t(\theta_*)$ and $g_t(\gamma_*)$ for $f(Y_t|Z_t:\theta_*)$ and $g(Y_t|Z_t;\gamma_*)$ respectively, and where:

$$\bar{V}_n = \begin{bmatrix} \bar{V}_{\theta\theta n} & ; & \bar{V}_{\theta\gamma n} \\ \bar{V}_{\gamma\theta n} & : & \bar{V}_{\gamma\gamma n} \end{bmatrix},$$

$$\bar{V}_{\theta\theta n} = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial \log f_t(\bar{\theta}_n)}{\partial\theta} \cdot \frac{\partial \log f_t(\bar{\theta}_n)}{\partial\theta'} + \frac{1}{n}\sum_{t=1}^{n}\left[\log \frac{f_t(\bar{\theta}_n)}{g_t(\bar{\gamma}_n)}\right]\frac{\partial^2 \log f_t(\bar{\theta}_n)}{\partial\theta\partial\theta'},$$

$$\bar{V}_{\theta\gamma n} = \bar{V}_{\gamma\theta n}' = -\frac{1}{n}\sum_{t=1}^{n} \frac{\partial \log f_t(\bar{\theta}_n)}{\partial\theta} \cdot \frac{\partial \log g_t(\bar{\gamma}_n)}{\partial\gamma'},$$

$$\bar{V}_{\gamma\gamma n} = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial \log g_t(\bar{\gamma}_n)}{\partial\gamma} \cdot \frac{\partial \log g_t(\bar{\gamma}_n)}{\partial\gamma'} - \frac{1}{n}\sum_{t=1}^{n}\left[\log \frac{f_t(\bar{\theta}_n)}{g_t(\bar{\gamma}_n)}\right]\frac{\partial^2 \log g_t(\bar{\gamma}_n)}{\partial\gamma\partial\gamma'},$$

for some $\bar{\theta}_n$ and $\bar{\gamma}_n$ in the segments $[\theta_*,\hat{\theta}_n]$ and $[\gamma_*,\hat{\gamma}_n]$ respectively.

But, $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ under H_0^ω (Lemma 4.1) so that the first three terms in (A.19) are null. Moreover, given Assumption A1-A7, Jennrich's uniform strong Law of Large Numbers, the second term in $\bar{V}_{\theta\theta n}$ (or $\bar{V}_{\gamma\gamma n}$) goes almost surely to zero since $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ under H_0^ω. Hence $\bar{V}_{\theta\theta n} = B_f + o_p(1)$, $\bar{V}_{\gamma\gamma n} = B_g + o_p(1)$, $\bar{V}_{\theta\gamma n} = \bar{V}_{\gamma\theta n}' = -B_{fg} + o_p(1)$. Since $n^{1/2}(\hat{\theta}_n - \theta_*)$ and $n^{1/2}(\hat{\gamma}_n - \gamma_*)$ are both $O_p(1)$, it follows that under H_o^ω:

$$n\hat{\omega}_n^2 = n\tilde{\omega}_n^2 + o_p(1)$$

$$= n(\hat{\theta}_n' - \theta_*', \hat{\gamma}_n - \gamma_*')V(\hat{\theta}_n' - \theta_*', \hat{\gamma}_n' - \gamma_*')' + o_p(1) \tag{A.20}$$

where

$$V = \begin{bmatrix} B_f & -B_{fg} \\ -B_{gf} & B_g \end{bmatrix}. \tag{A.21}$$

From Lemmae 2.1 and 3.4, it remains to show that the eigenvalues of $V\Sigma$ (or $\Sigma^{1/2}V\Sigma^{1/2}$) are equal to the squares of the eigenvalues of $W = Q\Sigma$ (or $\Sigma^{1/2}Q\Sigma^{1/2}$) where Q is defined in Equation (A.11). It is easy to check that $V = Q\Sigma Q$. Hence if R is the matrix that orthogonalizes $\Sigma^{1/2}Q\Sigma^{1/2}$ so that $R\Sigma^{1/2}Q\Sigma^{1/2}R' = \Lambda_*$, then $\Lambda_*^2 = R\Sigma^{1/2}Q\Sigma Q\Sigma^{1/2}R' = R\Sigma^{1/2}V\Sigma^{1/2}R'$. This completes the proof.

<u>Proof of Theorem 4.4</u>: From Lemma 2.1, Equation (A.20), and Rao and Mitra (1971, Theorem 9.2.1), it follows that $n\hat{\omega}_n^2$ (or $n\tilde{\omega}_n^2$) has a limiting (central) chi- square distribution if and only if $\Sigma V\Sigma V\Sigma = \Sigma V\Sigma$ in which case the number of degrees of freedom is tr $V\Sigma$. Using the information matrix equivalences (3.11), we have:

$$V\Sigma = \begin{bmatrix} I_p - B_{fg}B_g^{-1}B_{gf}B_f^{-1} & ; & 0 \\ 0 & ; & I_q - B_{gf}B_f^{-1}B_{fg}B_g^{-1} \end{bmatrix}. \tag{A.22}$$

$$\Sigma V\Sigma = \begin{bmatrix} B_f^{-1}(I_p - B_{fg}B_g^{-1}B_{gf}B_f^{-1}) & ; & B_f^{-1}B_{fg}B_g^{-1}(I_q - B_{gf}B_f^{-1}B_{fg}B_g^{-1}) \\ B_g^{-1}B_{gf}B_f^{-1}(I_p - B_{fg}B_g^{-1}B_{gf}B_f^{-1}) & ; & B_g^{-1}(I_q - B_{gf}B_f^{-1}B_{fg}B_g^{-1}) \end{bmatrix},$$

$$\Sigma V\Sigma V\Sigma = \begin{bmatrix} B_f^{-1}(I_p - B_{fg}B_g^{-1}B_{gf}B_f^{-1})^2 & ; & B_f^{-1}B_{fg}B_g^{-1}(I_q - B_{gf}B_f^{-1}B_{fg}B_g^{-1})^2 \\ B_g^{-1}B_{gf}B_f^{-1}(I_p - B_{fg}B_g^{-1}B_{gf}B_f^{-1})^2 & ; & B_g^{-1}(I_q - B_{gf}B_f^{-1}B_{fg}B_g^{-1})^2 \end{bmatrix}.$$

Hence $\Sigma V\Sigma V\Sigma = \Sigma V\Sigma$ if and only if $I_p - B_{fg}B_g^{-1}B_{gf}B_f^{-1}$ and $I_q - B_{gf}B_f^{-1}B_{fg}B_g^{-1}$ are <u>both</u> idempotent. Or equivalently $\Sigma V\Sigma V\Sigma = \Sigma V\Sigma$ if and only if $B_{fg}B_g^{-1}B_{gf}B_f^{-1}$ and $B_{gf}B_f^{-1}B_{fg}B_g^{-1}$ are <u>both</u> idempotent.

But, $B_{fg}B_g^{-1}B_{gf}B_f^{-1}$ is idempotent if and only if $B_{gf}B_f^{-1}B_{fg}B_g^{-1}$ is idempotent. Indeed, rank $(B_{fg}B_g^{-1})(B_{gf}B_f^{-1})$ = rank $B_{fg}B_g^{-1}B_{gf}$ = rank B_{fg} = rank $B_{fg}B_g^{-1}$. Thus, from Rao and Mitra (1971, Lemma 2.2.7), it follows that if $(B_{fg}B_g^{-1})(B_{gf}B_f^{-1})$ is idempotent then $(B_{gf}B_f^{-1})(B_{fg}B_g^{-1})$ is also idempotent. By the same argument, if $B_{gf}B_f^{-1}B_{fg}B_g^{-1}$ is idempotent then $B_{fg}B_g^{-1}B_{gf}B_f^{-1}$ is also idempotent. This establishes the equivalence between (i), (ii), (iii), and (iv). Finally, from (A.22):

$$\text{tr } V\Sigma = p + q - \text{tr}(B_{fg}B_g^{-1}B_{gf}B_f^{-1}) - \text{tr}(B_{gf}B_f^{-1}B_{fg}B_g^{-1}),$$

$$= p + q - 2\text{tr}(B_{fg}B_g^{-1}B_{gf}B_f^{-1}).$$

Since $B_{fg}B_g^{-1}B_{gf}B_f^{-1}$ must be idempotent for $n\hat{\omega}_n^2$ to be chi-square distributed asymptotically, then $\text{tr}(B_{fg}B_g^{-1}B_{gf}B_f^{-1})$ = rank $(B_{fg}B_g^{-1}B_{gf}B_f^{-1})$ = rank B_{gf}. This establishes that the number of degrees of freedom is $p + q - 2$ rank B_{gf}.

<u>Proof of Theorem 5.2</u>: Straightforward from Theorem 3.5 - (ii), and Lemma 4.2 since $f(\cdot|\cdot;\theta_*) \neq g(\cdot|\cdot;\gamma_*)$ and $\omega_*^2 > 0$.

<u>Proof of Corollary 5.3</u>: Obvious from Equation (5.13) and Theorem 5.2.

<u>Proof of Theorem 5.4</u>: To prove Part (i), note that under $\tilde{H}_o$: $\Delta = 0$ so that by subtractiving $n\Delta$ from $L\tilde{R}_n(\hat{\theta}_n,\hat{\gamma}_n)$ we obtain after multiplication by $n^{-1/2}/\hat{\omega}_n$:

$$\frac{n^{-1/2}}{\hat{\omega}_n}L\tilde{R}_n(\hat{\theta}_n,\hat{\gamma}_n) = \frac{1}{\hat{\omega}_n}[n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n) - n^{1/2}E^o[\log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}]].$$

Since $f(\cdot|\cdot;\theta_*) \neq g(\cdot|\cdot;\gamma_*)$ because the models are strictly non-nested, Part (i) follows from Theorem 3.5 - (ii) and Lemma 4.2 - (i).

To prove Parts (ii) and (iii), note that

$$\frac{n^{-1/2}}{\hat{\omega}_n}\widetilde{LR}_n(\hat{\theta}_n,\hat{\gamma}_n) = \frac{1}{\hat{\omega}_n}\left[n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n) - n^{1/2}E^0\left[\log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right]\right] + \frac{n^{1/2}}{\hat{\omega}_n}\Lambda.$$

The first term is $O_p(1)$ from Theorem 3.5 - (ii), and the second term goes almost surely to $+\infty$ under $\tilde{H}_f$ and to $-\infty$ under $\tilde{H}_g$.

<u>Proof of Theorem 6.2</u>: Part (i) follows from Theorem 4.3, since the c.d.f. $M_{p+q}(\cdot;\lambda)$ is continuous in λ, and since $\hat{W}_n$ converges almost surely to W so that the eigenvalues $\hat{\lambda}_n$ converge also almost surely to λ_*. Part (ii) follows from Lemma 4.2 - (i). Part (iii) follows by the same argument.

<u>Proof of Lemma 6.3</u>: We shall prove that (ii) $\Rightarrow$ (i) $\Rightarrow$ (iii) $\Rightarrow$ (ii). Without loss of generality, we assume that $H^0(\cdot|\cdot) \in F_\theta$, i.e., that $h^0(\cdot|\cdot) = f(\cdot|\cdot;\theta_0)$ for some θ_0 in Θ. Then, as is well known, it follows from the uniqueness of θ_* (Assumption A3 - (b)) and Jensen's inequality that $\theta_* = \theta_0$. Thus $h^0(\cdot|\cdot) = f(\cdot|\cdot;\theta_*)$

(ii) $\Rightarrow$ (i): Since $h^0(\cdot|\cdot) = f(\cdot|\cdot;\theta_*)$, then $h^0(\cdot|\cdot) = g(\cdot|\cdot;\gamma_*)$ using (ii), so that $H^0(\cdot|\cdot) \in G_\gamma$, and hence $H^0(\cdot|\cdot) \in F_\theta \cap G_\gamma$.

(i) $\Rightarrow$ (iii): Since $H^0(\cdot|\cdot) \in G_\gamma$, then $h^0(\cdot|\cdot) = g(\cdot|\cdot;\gamma_*)$ as above. Since $h^0(\cdot|\cdot) = f(\cdot|\cdot;\theta_*)$, then $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$, which implies (iii).

(iii) $\Rightarrow$ (ii): Since $h^0(\cdot|\cdot) = f(\cdot|\cdot;\theta_*)$, then (iii) implies that:

$$\int_Z\{\int_Y \log\frac{f(y|z;\theta_*)}{g(y|z;\theta_*)} f(y|z;\theta_*)dy\}dz = 0.$$

Then (ii) follows from Jensen's inequality.

<u>Proof of Theorem 6.4</u>: Under H_0, it follows from Lemma 6.3 that $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$. Then, Part (i) follows from Theorem 3.4 - (i), the continuity of the c.d.f. $M_{p+q}(\cdot;\lambda)$ in λ, and the strong convergence of $\hat{\lambda}_n$ to the eigenvalues λ_* of W. Parts (ii) and (iii) follow from Lemma 3.1.

<u>Proof of Lemma 7.2</u>: We shall prove that (ii) $\Rightarrow$ (i) $\Rightarrow$ (iv) $\Rightarrow$ (iii) $\Rightarrow$ (ii).

(ii) $\Rightarrow$ (i): Since $\theta_* \in \phi(\Gamma)$, $\exists\tilde{\gamma} \in \Gamma$ such that $\theta_* = \phi(\tilde{\gamma})$. Thus, from Assumption A8, $\log g(\cdot|\cdot;\tilde{\gamma}) = \log f(\cdot|\cdot;\theta_*)$ which implies $E^0[\log g(Y_t|Z_t;\tilde{\gamma})] = E^0[\log f(Y_t|Z_t;\theta_*)] \geq E^0[\log f(Y_t|Z_t;\theta)]$ for any θ in Θ and, in particular for any θ in $\phi(\Gamma)$, i.e., for any $\theta = \phi(\gamma)$ for $\gamma \in \Gamma$. Then, using again Assumption A8, we have $E^0[\log g(Y_t|Z_t;\tilde{\gamma})] \geq E^0[\log g(Y_t|Z_t;\gamma)]$ for any $\gamma \in \Gamma$, which implies that $\tilde{\gamma} = \gamma_*$ from Assumption A3 - (b), and hence that $\theta_* = \phi(\gamma_*)$.

(i) $\Rightarrow$ (iv): Obvious given Assumption A8.

(iv) $\Rightarrow$ (iii): Obvious.

(iii) $\Rightarrow$ (ii): Suppose that $\theta_* \notin \phi(\Gamma)$, then $\theta_* \neq \tilde{\theta} \equiv \phi(\gamma_*)$. But from (iii) and Assumption A8, we have $E^0[\log f(Y_t|Z_t;\theta_*)] = E^0[\log f(Y_t|Z_t;\tilde{\theta})]$, which contradicts the uniqueness of θ_* (Assumption

A3 - (b)).

<u>Proof of Lemma 7.3</u>: First, we note that, under Assumption A8, $\partial \log g(\cdot|\cdot;\gamma)/\partial\gamma = \partial d'/\partial\gamma \cdot \partial \log f(\cdot|\cdot;d(\gamma))/\partial\theta$. But under H_0^θ, we have $\theta_* = d(\gamma_*)$ (Lemma 7.2), which establishes Part (ii) and the first equality of Part (i) using the definitions of B_g, B_f, and B_{gf}. In addition:

$$\frac{\partial^2 \log g}{\partial\gamma\partial\gamma'} = \frac{\partial d'}{\partial\gamma} \cdot \frac{\partial^2 \log f}{\partial\theta\partial\theta'} \cdot \frac{\partial d}{\partial\gamma'} + \sum_k \frac{\partial d_k}{\partial\gamma\partial\gamma'} \frac{\partial \log f}{\partial\theta_k},$$

where we have omitted the arguments of the functions, and where d_k is the k-th component of d. Since $E^0[\partial \log f(Y_t|Z_t;\theta_*)/\partial\theta] = 0$ and since $\theta_* = d(\gamma_*)$, then the second equality of Part (i) follows. Finally, Part (iii) follows from this equality and the fact that $A_f(\theta_*)$ and $A_g(\gamma_*)$ are non- singular matrices (see, White (1982a), Theorem 3.1)).

<u>Proof of Theorem 7.4</u>: Since under H_0^θ, we have $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ (Lemma 7.2), then Part (i) follows from Theorem 3.5 - (i) if we show that the non-zero eigenvalues λ_* of W are the non-zero eigenvalues of $\underline{W}$. But, using Lemma 7.3, the eigenvalues of W solve:

$$0 = \det \begin{bmatrix} -B_f A_f^{-1} - \lambda I_p & ; & -B_f \frac{\partial d}{\partial\gamma'} A_g^{-1} \\ \frac{\partial d'}{\partial\gamma} B_f A_f^{-1} & ; & \frac{\partial d'}{\partial\gamma} B_f \frac{\partial d}{\partial\gamma'} A_g^{-1} - \lambda I_q \end{bmatrix},$$

$$= \det \begin{bmatrix} -B_f A_f^{-1} - \lambda I_p & ; & -B_f \frac{\partial d}{\partial\gamma'} A_g^{-1} \\ -\lambda \frac{\partial d'}{\partial\gamma} & ; & -\lambda I_q \end{bmatrix},$$

$$= \det \begin{bmatrix} -B_f A_f^{-1} - \lambda I_p + B_f \frac{\partial d}{\partial\gamma'} A_g^{-1} \frac{\partial d'}{\partial\gamma} & ; & -B_f \frac{\partial d}{\partial\gamma'} A_g^{-1} \\ 0 & ; & -\lambda I_q \end{bmatrix},$$

where the second equation follows from the first equation by adding to the second-row matrices the first-row matrices premultiplied by the full row-rank matrix $\partial d'/\partial\gamma$ (Lemma 7.3 - (iii)), and where the third equation follows from the second equation by adding to the first-column matrices the second-column matrices postmultiplied by $-\partial d'/\partial\gamma$. Hence, the eigenvalues of W solve:

$$0 = \lambda^q \det\{-B_f A_f^{-1} + B_f \frac{\partial d}{\partial\gamma'} A_g^{-1} \frac{\partial d'}{\partial\gamma} - \lambda I_p\}, \tag{A.23}$$

which establishes that the non-zero eigenvalues of W are the non-zero eigenvalues of $\underline{W}$ as defined by Equation (7.5). Equation (A.23) also shows that the eigenvalues of $\underline{W}$ are all real and non-negative since $A_f^{-1} - [\partial d/\partial\gamma']A_g^{-1}[\partial d'/\partial\gamma] = A_f^{-1} - [\partial d/\partial\gamma']([\partial d'/\partial\gamma]A_f[\partial d/\partial\gamma'])^{-1}[\partial d'/\partial\gamma]$ which is n.s.d.

Part (ii) follows from Lemma 3.1 and $H_A^\theta = H_f$.

<u>Proof of Corollary 7.5</u>: If $A_f + B_f = 0$, then it follows from Lemma 7.3 - (i) that under H_0^θ, $A_g + B_g = 0$. Part (i) follows from Theorem 3.6 and Lemma 7.3 since Condition (3.12) is satisfied. Part (ii) is identical to Theorem 7.4 - (ii).

<u>Proof of Theorem 7.6</u>: Since $H_o^\theta = H_o^\omega$, Part (i) follows from Theorem 4.3 since the non-zero eigenvalues of W are the eigenvalues of $\underline{W}$ (see

the proof of Theorem 7.4). Parts (ii) follows from Lemma 4.2 since H_A^θ is equivalent to H_A^ω. Part (iii) is proved similarly.

<u>Proof of Corollary 7.7</u>: As noticed in the proof of Corollary 7.5, given the assumptions of Corollary 7.7, we have both information matrix equivalences (3.11) under H_0^θ. Then Part (i) follows from Theorem 4.4 - (iv) by noticing that the matrix $B_{gf}B_f^{-1}B_{fg}B_g^{-1}$ is equal to I_q (using Lemma 7.3) and hence is idempotent. Parts (ii) and (iii) are identical to Parts (ii) and (iii) of Theorem 7.6.

FOOTNOTES

* This research was supported by National Science Foundation Grant SES-8410593. I am indebted to P. Bjorn, D. Lien, D. Rivers for helpful discussions, and to J. M. Dufour for some references on weighted sums of chi-square distributions. I would like to thank especially H. White whose comments much improved this paper. I am also grateful to C. R. Jackson without whom this paper would not have been written and to L. Donnelly for stimulating thoughts. Remaining errors are mine.

1. The notation $o_p(1)$ indicates a quantity that converges in probability to zero, while the notation $O_p(1)$ indicates a quantity that is bounded in probability as n goes to infinity (see, e.g., Mann and Wald (1943)). As a matter of fact, Equation (3.4) holds whether or not $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$. The point is that, if $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$, then the asymptotic distribution of the LR statistic will be given by Equation (3.3).

2. As noticed earlier, only the $\underline{m}$ non-zero eigenvalues $\underline{\lambda}$ are relevant, i.e., $M_m(\cdot;\lambda) = M_{\underline{m}}(\cdot;\underline{\lambda})$. Moreover, these eigenvalues are all real, and that they are all non-negative if Q is positive semi-definite.

3. Since rank W = rank $\sum$ = r, then the limiting distribution in (3.8) is equal to $M_r(\cdot;\underline{\lambda}_*)$ where $\underline{\lambda}_*$ is the vector of non-zero eigenvalues of W. Let us also note that some eigenvalues λ_* may be negative since the matrix defining the quadratic form in Equation (3.3) is not p.s.d. (see footnote 2).

4. In fact, Property (3.10) holds whether or not $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ (see also footnote 1). However, $\omega_*^2 = 0$ if and only if $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ (see Lemma 4.1 below). Thus, one must instead rely on the asymptotic approximation (3.8).

5. Given the definition of θ_*, it is clear that $\mathrm{KLIC}(H^0_{Y|Z};F_\theta)$ is the minimum distance between the true conditional distribution $H^0(\cdot|\cdot)$ and any conditional distribution $F(\cdot|\cdot;\theta)$ in F_θ.

6. The case when F_θ and G_γ are not nested but do have a non-empty intersection is treated in the next section on overlapping models. for a long time, non-nested hypotheses were defined as hypotheses that cannot be obtained from the other by a suitable limiting approximation (Cox (1961, 1962)). Noting that there were no satisfactory definitions of this concept, Pesaran (1985) recently proposed formal definitions of globally non-nested, partially non-nested, and nested hypotheses based on the KLIC. It can be shown that Pesaran's definitions are equivalent to our Definitions 5.1, 6.1, and 7.1. Our definitions appear to be more intuitive and natural.

7. Note that, from Equation (4.4), it follows that:

$$n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\tilde{\omega}_n \leq n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n.$$

Thus, even though under the Pittman approach the tests will have the same asymptotic power, this inequality suggests that the test based on $\hat{\omega}_n$ will be asymptotically more powerful than the test based on $\tilde{\omega}_n$ according to other definitions of asymptotic power such as Bahadur (1960)'s definition.

8. I owe this point to Hal White.

9. The reason for this multitude of criteria is that Sawa (1978) and Chow (1981) question the validity of Akaike's initial derivation.

10. Note that a correctly specified model is no longer necessarily best. More generally, k(p) may depend on n.

11. In the univariate dichotomous case, Cox (1970) points out that the logit and probit models are approximations of each other. If the explanatory variables are all discrete, Lee (1981) points out that in the bivariate dichotomous case the probit and logit models are either identical or nested. Morimune (1979) proposes some Cox-type tests for discriminating between the logit and the probit models. As argued in Section 8, these tests are conceptually different from the ones proposed here.

12. The variance test can be avoided by testing only some implications of the hypothesis $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$. This is done by first characterizing the conditions that θ and γ must satisfy for $f(\cdot|\cdot;\theta)$ to be equal to $g(\cdot|\cdot;\gamma)$. (See Lien and Vuong (1986) for an illustration.) In general, tests of some appropriately selected conditions are easier to perform than the variance test, and can be done using only $\hat{\theta}_n$ or $\hat{\gamma}_n$. The difficulty is to derive these conditions.

13. To see this, Note that H_o is a composite of H_o^ω and $H_o - H_o^\omega$. Let $A \equiv \{n\omega_n^2 > c_1\}$ and $B \equiv \{|n^{-1/2}LR(\hat{\theta}_n,\hat{\gamma}_n)/\hat{\omega}_n| > c_2\}$. Then $\Pr[\text{reject } H_o|H_o] = \Pr[A \cap B|H_o] =$

$\max\{\Pr(A \cap B|H_o^{\omega}), \Pr(A \cap B|H_o - H_o^{\omega})\} \leq$
$\max\{\Pr(A|H_o^{\omega}), \Pr(B|H_o - H_o^{\omega})\}$. But from Theorems 5.2 and 6.2, $\Pr(A|H_o^{\omega}) \to \alpha_1$ and $\Pr(B|H_o - H_o^{\omega}) \to \alpha_2$.

14. Johnson and Kotz (1969) give values of $M_m(x;\lambda)$ for $m = 4$ and some values of x and λ with a Fortran IV program for calculating $M_m(x;\lambda)$ which can also be used to compute the upper-tail probability $1 - M_{p+q}(n\hat{\omega}_n^2;\hat{\lambda}_n^2)$. Paul Bjorn told me, however, that there are some problems with this Fortran program.

15. It follows that $H_0^{\omega} = H_0$ and $H_A^{\omega} = H_f \cup H_g$. The variance test of Theorem 6.2 can therefore be thought of as a discrimination test since the null and alternative hypotheses correspond respectively to the equivalence and non-equivalence of the models. Contrary to the LR- based test proposed below, the variance test is not directional in the sense that when one rejects H_0, one does not know if it is in favor of H_f or H_g.

16. It is worth noting that if one rejects H_0 using the LR-based test, then one knows if it is in favor of H_f or H_g. Since it is assumed that at least one model is correctly specified, then rejection in favor of H_f will imply that F_θ is correctly specified and G_γ is incorrectly specified. A similar research applies in case of rejection in favor of H_g.

17. In fact, the computation of both c_1 and c_2 can be replaced by the computation of only the upper-tail probability of $2LR_n(\hat{\theta}_n, \hat{\gamma}_n)$ from the distribution $M_{p+q}(\cdot;\hat{\lambda}_n)$.

18. It is assumed throughout this and the next sections that the information matrix equivalence holds whenever the model is correctly specified. This actually requires a mild additional assumption (see, e.g. White (1982, Assumption A7), Vuong (1983, Assumption A6)).

19. Classical nested hypothesis testing actually assumes that the larger model is correctly specified. Only recently this framework has been extended to the misspecified case (see, e.g., White (1982a)). Let us also note that the equivalence between model selection tests and nested hypothesis tests does not hold if one introduces a correction factor as in the criterion (5.14).

20. Though the variance tests are asymptotically equivalent, they are not asymptotically equivalent to the LR test under H_o^{θ}. In addition, these tests are not asymptotically equivalent under H_o^{θ} to the robust Wald and LM tests proposed by White (1982a) for testing the parametric restrictions H_o^{θ}. The relative asymptotic power properties of all these tests of H_o^{θ} in the misspecified case is left for future research.

21. Rejection of the equivalence in favor of the competing model being better does not, of course, imply that the alternative model is correctly specified. Note also that rejection of the equivalence in favor of the initial model being better implies that the alternative model is misspecified.

22. For subsequent work on specification tests, see Newey (1983), Ruud (1984), Vuong (1983,1984), among others.

23. Another important difference is that most specification tests use

only estimators of the model under test while our model selection tests use estimators of both the initial model and the competing model. This difference is similar to the one between the score or lagrange multiplier test and the LR test in the familiar nested hypothesis framework. Though the current specification tests and our LR based tests may have identical local power properties, our model selection tests are likely to have nicer global power properties (see, e.g., Bahadur (1967)).

24. See White (1982b) and Aguirre-Torres and Gallant (1983). White (1982b) showed that the test based on T_n^f is asymptotically equivalent to the J and P tests proposed by Davidson and MacKinnon (1981). Originally, Cox (1961, 1962) used a different but asymptotically equivalent statistic which is given by Equation (8.1) where $g(y|Z_t;\hat{\gamma}_n)$ is replaced by $g(y|Z_t;\gamma_*(\hat{\theta}_n))$. This latter statistic was used by Pesaran (1974) and Pesaran and Deaton (1978). It is clear that Equation (8.2) still holds. Another Cox-type test for non- nested hypotheses is the one proposed by Morimune (1983).

25. See also Datsoor (1981) who observes that nT_n^f and nT_n^g cannot both go in probability to $+\infty$ in the linear regression context.

26. For a similar point of view in a Bayesian framework, see also Klein (1983). But, see also McKinnon (1983b).

27. Throughout, we assume that the support of $H^0_{y|z}$ is Y for H^0_z-almost all z. Then the open set $N_z = \{y; f(y|z;\theta_*) \neq Kg(y|z;\theta_*)\}$ must be empty. The result follows by integrating with respect to ν_y.

REFERENCES

Aguirre-Torres, V., and A. R. Gallant: "The Null and Non-Null Asymptotic Distribution of the Cox Test for Multivariate Nonlinear Regression: Alternatives and a New Distribution-Free Cox Test," Journal of Econometrics, 21(1983), 5-33.

Aitchinson, J., and S. D. Silvey: "Maximum Likelihood Estimation of Paramaters Subject to Restraints," Annals of Mathematical Statistics, 29(1958), 813-828.

Akaike, H.: "Information Theory and an Extension of the Likelihood Ratio Principle," Proceedings of the Second International Symposium of Information Theory, ed. by B. N. Petrov and F. Csaki. Budapest: Akademiai Kiado, 1973, 257-281.

__________: "A New Look at the Statistical Model Identification," IEEE Transactions on Automatic Control, AC-19(1974), 716-723.

Amemiya, T.: "Selection of Regressors," International Economic Review, 21(1980), 331-354.

Atkinson, A. C.: "A Test for Discriminating between Models," Biometrika, 56(1969), 337-347.

__________: "A Method for Discrimating between Models," Journal of the Royal Statistical Society, Series B, 32(1970), 323-353.

Bahadur, R. R.: "Stochastic Comparison of Tests," Annals of Mathematical Statistics, 31(1960), 276-295.

__________: "An Optimal Property of the Likelihood Ratio Statistics," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics. Berkeley: University of California Press, 1967, Vol. 1, 13-26.

Box, G. E. P., and D. R. Cox: "An Analysis of Transformations," Journal of the Royal Statistical Society, Series B, 26(1964), 211-243.

Burguette, J., A. R. Gallant, and G. Souza: "On the Unification of the Asymptotic Theory of Nonlinear Econometric Models," Econometric Reviews, 1(1982), 151-190.

Chow, G.: "Selection of Econometric Models by the Information Criterion," Proceedings of the Econometric Society European Meeting, ed. E. G. Charatsis. Amsterdam: North Holland, 1981.

_________: Econometrics. New York: McGraw-Hill, 1983.

Cox, D. R.: "Tests of Seperate Families of Hypotheses," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1(1961), 105-123.

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

_________: The Analysis of Binary Data. London: Methuen, 1970.

Dastoor, N. K.: "A Note on the Interpretation of the Cox Procedure for Non-Nested Hypotheses," Economics Letters, 8(1981), 113-119.

Davidson, R., and J. G. MacKinnon: "Several Tests for Model Specification in the Presence of Alternative Hypotheses," Econometrica, 49(1981), 781-793.

Fisher, G., and M. McAleer: "On the Interpretation of the Cox Test in Econometrics," Economics Letters, 4(1979), 145-150.

Foutz, R. V., and R. C. Srivastana: "The Performance of the Likelihood Ratio Test When the Model is Incorrect," Annals of Statistics, 5(1977), 1183-1194.

Gallant, A. R., and A. Holly: "Statistical Inference in an Implicit, Nonlinear, Simultaneous Equation Model in the Context of Maximum Likelihood Estimation," Econometrica, 48(1980), 697-720.

Gourieroux, C., A. Monfort, and A. Trognon: "Testing Nested or Non-Nested Hypotheses," Journal of Econometrics, 21(1983), 83-115.

_________: "Pseudo Maximum Likelihood Method: Theory," Econometrica, 52 (1984), 681-700.

Hausman, J.: "Specification Tests in Econometrics," Econometrica 46(1978), 1251-1272.

Hendry, D. F.: "Comment," Econometric Reviews, 2(1983), 111-114.

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

Jennrich, R. I.: "Asymptotic Properties of Non-Linear Least Squares Estimators," Annals of Mathematical Statistics, 40(1969), 633-643.

Johnson, N. L., and S. Kotz: "Tables of Distributions of Positive Definite Quadratic Forms in Central Normal Variables," Sankhya, Series B, (1969), 303-314.

_________: Continuous Univariate Distributions - 2. New York: John Wiley and Sons, 1970.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, and T. C. Lee: The Theory and Practice of Econometrics. New York: John Wiley and Sons, Second edition, 1985.

Kent, J. T.: "Robust Properties of Likelihood Ratio Tests," Biometrika, 69(1982), 19-27.

Klein, R. W.: "Comment," Econometric Reviews, 2(1983), 115-119.

Kullback, S., and R. A. Leibler: "On Information and Sufficiency," Annals of Mathematical Statistics, 22(1951), 79-86.

Lee, L. F.: "On Comparisons of Normal and Logistic Models in the Bivariate Dichotomous Analysis," Economic Letters, 4(1979), 151- 155.

Lien, D., and Q. H. Vuong: "Selecting the Best Linear Regression Model: A Classical Approach," mimeo, California Institute of Technology, 1986.

MacKinnon, J. G.: "Model Specification Tests against Non-Nested Alternatives," Econometric Reviews, 2(1983), 85-110.

_________: "Reply," Econometric Reviews, 2(1983), 151-158.

Mann, H. B., and A. Wald: "On Stochastic Limit and Order Relationships," Annals of Mathematical Statistics, 14(1984), 217- 226.

McAleer, M., and A. Bera: "Comment," Econometric Reviews, 2(1983), 121-130.

Mizon, G., and J. F. Richard: "The Encompassing Principle and its Application to Non-Nested Hypotheses," Discussion Paper, University of Southampton, 1982.

Moore, D. S.: "Chi-Square Tests," in Studies in Statistics, R. V. Hogg (ed.), The Mathematical Association of America, Volume 19, 1978.

Morimune, K.: "Comparisons of Normal and Logistic Models in the Bivariate Dichotomous Analysis," Econometrica, 47(1979), 957-975.

__________: "Comment," Econometric Reviews, 2(1983), 137-143.

Newey, W.: "Maximum Likelihood Specification Testing and Instrumented Score Tests," mimeographed, Princeton University, 1983.

Neyman, J., and E. S. Pearson: "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," Biometrika, 20A(1928), 175-240.

Pesaran, M. H.: "On the General Problem of Model Selection," Review of Economic Studies," 41(1974), 153-171.

__________, "Global and Partial Non-nested Hypotheses and Asymptotic Local Power," mimeo, Trinity College, 1985.

Pesaran, M. H., and A. S. Deaton: "Testing Non-Nested Nonlinear Regression Models," Econometrica, 46(1978), 677-694.

Rao, C. R.: Linear Statistical Inference and its Applications. New York: John Wiley and Sons, 1973.

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

Ruud, P.: "Tests of Specification in Econometrics," Econometric Reviews, 3(1984), 211-242.

Sawa, T.: "Information Criteria for Discriminating among Alternative Regression Models," Econometrica, 46(1978), 1273-1291.

Schwarz, G.: "Estimating the Dimension of a Model," Annals of Statistics, 6(1978), 461-464.

Silvey, S. D.: "The Lagrangian Multiplier Test," Annals of Mathematical Statistics, 30(1959), 389-407.

Vuong, Q. H.: "Misspecification and Conditional Maximum Likelihood Estimation," Social Science Working Paper, No. 503. Pasadena: California Institute of Technology, 1983.

__________: "Two-Stage Conditional Maximum Likelihood Estimation of Econometric Models," Social Science Working Paper, No. 538. Pasadena: California Institute of Technology, 1984.

Wald, A.: "Tests of Statistical Hypotheses Concerning Several Parameters when the Number of Observations is Large," Transaction of the American Mathematical Society, 54(1943), 426-482.

White, H.: "Maximum Likelihood Estimation of Misspecified Models,"

Econometrica, 50(1982), 1-25.

__________: "Regularity Conditions for Cox's Test of Non-Nested Hypotheses," Journal of Econometrics, 19(1982), 301-318.

__________: "Editor's Introduction," Journal of Econometrics, 21(1983), 1-3.

White, H. and L. Olson: "Determinants of Wage Change on the Job: A Symmetric Test of Non-nested Hypotheses," mimeo, University of Rochester, 1979.

Wilks, S. S.: "The Large Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses," Annals of Mathematical Statistics, 9(1958), 60-62.