Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses

Author(s): Quang H. Vuong

Source: *Econometrica*, Mar., 1989, Vol. 57, No. 2 (Mar., 1989), pp. 307-333

Published by: The Econometric Society

Stable URL: https://www.jstor.org/stable/1912557

Econometrica, Vol. 57, No. 2 (March, 1989), 307–333

LIKELIHOOD RATIO TESTS FOR MODEL SELECTION AND NON-NESTED HYPOTHESES[1]

BY QUANG H. VUONG

In this paper, we develop a classical approach to model selection. Using the Kullback-Leibler Information Criterion to measure the closeness of a model to the truth, we propose simple likelihood-ratio based statistics for testing the null hypothesis that the competing models are equally close to the true data generating process against the alternative hypothesis that one model is closer. The tests are directional and are derived successively for the cases where the competing 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. We show that it is a weighted sum of chi-square distribution or a normal distribution depending on whether the distributions in the competing models closest to the truth are observationally identical. We also propose a test of this latter condition.

KEYWORDS: Likelihood ratio tests, model selection, non-nested hypotheses, misspecified models, weighted sums of chi-squares.

1. INTRODUCTION

THE MAIN PURPOSE OF THIS PAPER is to propose some new tests for model selection and non-nested hypotheses. 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, Ch. 9), Judge et al. (1985, Ch. 21)), we 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). See also Chow (1980). 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 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

[1]This research was supported by National Science Foundation Grant SES-8410593. An early version was presented at the North American Econometric Society meeting, New Orleans, 1986. I am indebted to P. Bjorn, D. Lien, D. Rivers, the co-editor, two referees, and seminar participants at the University of Southern California, University of California-Berkeley, Stanford University, University of Minnesota, University of Wisconsin, Yale University, MIT/Harvard University, University of Pennsylvania, University of Florida-Gainesville, North Carolina State/Duke University, Indiana University, and University of California-Irvine. I would like to thank especially H. White whose comments much improved this paper. I am grateful to C. R. Jackson and to L. Donnelly for stimulating thoughts. Remaining errors are mine. This paper is dedicated to some of my former colleagues at Caltech.

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, Rule 2.1)).

We consider conditional models so as to allow for explanatory variables. Then, if $\boldsymbol{F}_\theta = \{ f(y|z; \theta); \theta \in \Theta \}$ 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 $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma = \{ g(y|z; \gamma); \gamma \in \Gamma \}$, which may be nested, non-nested, or overlapping, we 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 $\boldsymbol{F}_\theta$ is better than $\boldsymbol{G}_\gamma$ or against $E^0[\log f(y|z; \theta_*)] < E^0[\log g(y|z; \gamma_*)]$ meaning that $\boldsymbol{G}_\gamma$ is better than $\boldsymbol{F}_\theta$. Tests of such hypotheses are called *tests for model selection*. Since the true density $h^0(y|z)$ is not restricted a priori to belong to either one of the models $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$, by necessity, the concern of this paper is 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 $E^0[\log f(y|z; \theta_*)] - E^0[\log g(y|z; \gamma_*)]$. 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, it is necessary to obtain the asymptotic distribution of the LR statistic under the most general conditions. To do so, we use the 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 noncentral chi-square distribution under sequences of local alternatives (Wald (1943)). However, as Foutz and Srivastava (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 redefined in terms of the pseudo-true values satisfying the specified restrictions. Parallel to this literature on 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. These results suggest that the asymptotic distribution of the LR statistic as well as the speed at which it converges to that distribution depend on whether the models are nested or correctly specified. In the first part of this paper, we completely characterize the asymptotic distribution of the LR statistic under the most general conditions.

The paper is organized as follows. In Section 2, we present the basic framework. In Section 3, we derive the asymptotic distribution of the LR statistic whether or not the models are nested or misspecified. We show that it depends on the condition $f(y|z;\theta_*) = g(y|z;\gamma_*)$ for almost all (y, z). In Section 4, we show that $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ is equivalent to the hypothesis that a variance ω_*^2 is zero. This allows us to construct a test of $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ based on a consistent estimator $\hat{\omega}_n^2$ of ω_*^2. In the next three sections, we apply the previous results to derive new and directional LR based tests for model selection in all possible situations. In Sections 5, 6, and 7, we consider successively the cases where the competing models are (strictly) non-nested, overlapping, and nested. We also briefly compare our approaches to that of Akaike (1973, 1974) and Cox (1961, 1962). Section 8 summarizes our results and suggests some directions for further research. All the proofs are collected in the Appendix.

2. BASIC FRAMEWORK

Let X_t be a $m \times 1$ observed random vector taking its values in a Polish space $\boldsymbol{X}$, i.e., a complete separable metric space. For instance, in the case of a real random vector, $\boldsymbol{X}$ is the Euclidean space $\mathbb{R}^m$. Let σ_X be the Borel σ-algebra on $\boldsymbol{X}$. 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 $m = l + k$. Let $(\boldsymbol{Y}, \sigma_Y)$ and $(\boldsymbol{Z}, \sigma_Z)$ be the measurable spaces associated with Y_t and Z_t. Let H_X^0 be the true joint distribution of X_t. We shall be interested in the true conditional distribution $H_{Y|Z}^0(\cdot|\cdot)$ of Y_t given Z_t, which exists by Jirina's Theorem (see, e.g., Bauer (1972, p. 319), Monfort (1980, p. 93)). We can think of Y_t as being the endogenous variables, and of Z_t as being the exogenous variables.

The process generating the observations X_t, $t = 1, 2, \ldots$, satisfies the next assumption. Let H_Z^0 be the true marginal distribution of Z_t, and ν_Y be a σ-finite measure on $(\boldsymbol{Y}, \sigma_Y)$.

ASSUMPTION A1: (a) *The random vectors* X_t, $t = 1, 2, \ldots$, *are independent and identically distributed* (*i.i.d.*) *with common true distribution* H_X^0 *on* $(\boldsymbol{X}, \sigma_X)$. (b) *For* H_Z^0*-almost all* z, $H_{Y|Z}^0(\cdot|z)$ *has a Radon-Nikodym density* $h^0(\cdot|z)$ *relative to* ν_Y, *which is strictly positive for* ν_Y*-almost all* y.

Assumption A1-(a) is more suitable for cross-section than time-series data. Some of our results can be generalized to more general data generating processes such as those considered by Burguette, Gallant, and Souza (1982), and White and Domowitz (1984). An assumption equivalent to Assumption A1-(b) is that

$H^0_{Y|Z}(\cdot\,|z)$ and ν_Y are, for H^0_Z-almost all z, absolutely continuous relative to each other (see, e.g., Bauer (1972, p. 901)). Since a similar remark applies to Assumption A2-(a) below, it follows that the measures $H^0_{Y|Z}(\cdot\,|z)$, $F_{Y|Z}(\cdot\,|z;\theta)$, and ν_Y are absolutely continuous relative to each other, and hence have the same negligible sets. As a consequence of these assumptions, the true conditional distribution $H^0_{Y|Z}(\cdot\,|\,\cdot)$ and the competing conditional models have the same support.[2]

We now consider two competing parametric families of conditional distributions defined on $\sigma_Y \times \boldsymbol{Z}$ for Y_t given Z_t: $\boldsymbol{F}_\theta \equiv \{F_{Y|Z}(\cdot\,|\,\cdot\,;\theta);\ \theta \in \Theta \subset \mathbb{R}^p\}$ and $\boldsymbol{G}_\gamma \equiv \{G_{Y|Z}(\cdot\,|\,\cdot\,;\gamma);\ \gamma \in \Gamma \subset \mathbb{R}^q\}$. No assumption is here made on the relationship between the two competing conditional models $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ 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 the following regularity conditions (Vuong (1983)) which are similar to those of White (1982a, 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 $\boldsymbol{F}_\theta$. It is understood that similar assumptions are made on $\boldsymbol{G}_\gamma$.

ASSUMPTION A2: (a) *For every θ in Θ and for H^0_Z-almost all z the conditional distribution $F_{Y|Z}(\cdot\,|z,\theta)$ has a Radon-Nikodym density $f(\cdot\,|z;\theta)$ relative to ν_Y, which is strictly positive for ν_Y-almost all y.* (b) *Θ is a compact subset of $\mathbb{R}^p$, and the conditional density $f(y|z;\theta)$ is continuous in θ for H^0_X-almost all (y, z).*

ASSUMPTION A3: (a) *For H^0_X-almost all (y, z), $|\log f(y|z;\cdot)|$ is dominated by an H^0_X-integrable function independent of θ.* (b) *The function $z_f(\theta) \equiv \int \log f(y|z;\theta) H^0_X(dx)$ has a unique maximum on Θ at θ_*.*

The value θ_* is called the pseudo-true value of θ for the conditional model $\boldsymbol{F}_\theta$ (see, e.g., Sawa (1978)). Similarly, γ_* denotes the pseudo-true value of γ for the conditional model $\boldsymbol{G}_\gamma$.

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

[2] Most of the results of this paper hold under the weaker assumption that ν_Y is absolutely continuous relative to $H^0_{Y|Z}(\cdot\,|z)$ for H^0_Z-almost all z. This latter assumption says that the non-negligible sets relative to ν_Y are also non-negligible relative to $H^0_{Y|Z}(\cdot\,|z)$. It does not require that $H^0_{Y|Z}(\cdot\,|z)$ have a density relative to ν_Y.

This ensures the existence of the usual matrices:

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

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

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 $\boldsymbol{G}_\gamma$. Moreover, since Assumption A4 holds for both models $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$, then for H_X^0-almost all (y, z), $|\partial \log f(y|z;\cdot)/\partial\theta \cdot \partial \log g(y|z;\cdot)/\partial y'|$ is dominated by an H_X^0-integrable function independent of θ and γ. This ensures the existence of the $p \times q$ matrix:

$$(2.3) \qquad B_{fg}(\theta,\gamma) = B'_{gf}(\gamma,\theta) \equiv E^0\left[\frac{\partial \log f(Y_t|Z_t;\theta)}{\partial\theta}\cdot\frac{\partial \log g(Y_t|Z_t;\gamma)}{\partial\gamma'}\right].$$

ASSUMPTION A5: (a) *θ_* is an interior point of Θ.* (b) *θ_* is a regular point of $A_f(\theta)$.*

Let σ_X^n be the n-product of σ_X. The (quasi) maximum likelihood (ML) estimator $\hat{\theta}_n$ for the conditional model $\boldsymbol{F}_\theta$ is a σ_X^n-measurable function of $(X_1, \ldots, X_n)$ such that

$$(2.4) \qquad L_n^f(\hat{\theta}_n) = \sup_{\theta\in\Theta} L_n^f(\theta),$$

where $L_n^f(\theta)$ is the conditional log-likelihood function for the model $\boldsymbol{F}_\theta$:

$$(2.5) \qquad L_n^f(\theta) \equiv \sum_{t=1}^{n} \log f(Y_t|Z_t;\theta).$$

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

$$(2.6) \qquad L_n^g(\gamma) \equiv \sum_{t=1}^{n} \log g(Y_t|Z_t;\gamma).$$

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 $A_f^{-1}(\theta_*)B_f(\theta_*)A_f^{-1}(\theta_*)$. Similar properties hold for the ML estimator $\hat{\gamma}_n$ of γ_*. As a matter of fact, $\hat{\theta}_n$ and $\hat{\gamma}_n$ are jointly asymptotically normal (see Lemma A in the Appendix) with asymptotic covariance matrix that can be consistently estimated using the sample analogs of $A_s(\theta)$, $B_s(\theta)$, and $B_{fg}(\theta,\gamma)$, $s = f, g$, evaluated at $(\hat{\theta}_n, \hat{\gamma}_n)$. For instance, $B_{fg}(\theta_*, \gamma_*)$

is consistently estimated by:

$$B_{fgn}(\hat{\theta}_n, \hat{\gamma}_n) = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial \log f(Y_t|Z_t; \hat{\theta}_n)}{\partial \theta} \cdot \frac{\partial \log g(Y_t|Z_t; \hat{\gamma}_n)}{\partial \gamma'}. \tag{2.7}$$

3. THE LIKELIHOOD RATIO STATISTIC

All our tests for model selection are based on the likelihood ratio (LR) statistic. In this section, we 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:

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

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

LEMMA 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 consider distributions of quadratic forms in normal random variables. Such distributions have been studied by, 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.

DEFINITION 1 (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 $\sum_{i=1}^{m} \lambda_i Z_i^2$ is distributed as a weighted sum of chi-squares with parameters (m, λ). Its cumulative distribution function (c.d.f.) is denoted by $M_m(\cdot; \lambda)$.

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 result allows Ω to be singular, and slightly differs from Moore (1978, Theorem 1).

LEMMA 3.2: *Let Y be a vector of m random variables distributed as $N(0, \Omega)$ with rank $\Omega \leqslant m$. Let Q be a $m \times m$ real symmetric matrix. Then*

$$Y'QY \sim M_m(\cdot; \lambda) \tag{3.3}$$

where λ is the vector of eigenvalues of $Q\Omega$. Moreover, the eigenvalues are all real, and they are all nonnegative if Q is positive semi-definite.

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_X^0 of (Y_t, Z_t). That is:

$$\omega_*^2 \equiv \operatorname{var}^0\left[\log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] = E^0\left[\log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right]^2 - \left[E^0\left[\log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right]\right]^2. \tag{3.4}$$

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

ASSUMPTION A6: *For H_X^0-almost all (y, z) the functions $|\log f(y|z;\cdot)|^2$ and $|\log g(y|z;\cdot)|^2$ are dominated by H_X^0-integrable functions independent of θ and γ.*

THEOREM 3.3 (Asymptotic Distribution of the LR Statistic): *Given Assumptions A1–A5:*

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

$$2LR_n(\hat\theta_n, \hat\gamma_n) \xrightarrow{D} M_{p+q}(\cdot;\lambda_*), \tag{3.5}$$

where λ_* is the vector of $p+q$ (possibly negative) eigenvalues of

$$W = \begin{bmatrix} -B_f(\theta_*)A_f^{-1}(\theta_*); & -B_{fg}(\theta_*,\gamma_*)A_g^{-1}(\gamma_*) \\ B_{gf}(\gamma_*,\theta_*)A_f^{-1}(\theta_*); & B_g(\gamma_*)A_g^{-1}(\gamma_*) \end{bmatrix}, \tag{3.6}$$

(ii) *if $f(\cdot|\cdot;\theta_*) \neq g(\cdot|\cdot;\gamma_*)$ and Assumption A6 holds, then*

$$n^{-1/2}LR_n(\hat\theta_n, \hat\gamma_n) - n^{1/2}E^0\left[\log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] \xrightarrow{D} N(0, \omega_*^2). \tag{3.7}$$

Throughout, the condition $f(\cdot|\cdot;\theta_*) = g(\cdot|\cdot;\gamma_*)$ is to be understood as holding for H_X^0-almost all (y, z), i.e., as $H_X^0[(y, z)\colon f(y|z;\theta_*) = g(y|z;\gamma_*)] = 1$. Its interpretation is that the distributions in $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ that are closest to the true conditional distribution $H_{Y|Z}^0(\cdot|\cdot)$ are observationally identical under H_X^0. Theorem 3.3 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 rate of convergence 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 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 assume that the information matrix equivalence holds for both $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$, i.e.:

$$A_f(\theta_*) + B_f(\theta_*) = 0 \quad \text{and} \quad A_g(\gamma_*) + B_g(\gamma_*) = 0. \tag{3.8}$$

As shown in White (1982a, Theorem 3.3), the information matrix equivalences hold under correct specification of the conditional models given mild additional assumptions.

COROLLARY 3.4 (Asymptotic Chi-Square Distribution of the LR Statistic given Information Matrix Equivalences): *Given Assumptions A1–A5, suppose that* (3.8) *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*:

$$B_g(\gamma_*) - B_{gf}(\gamma_*, \theta_*)B_f^{-1}(\theta_*)B_{fg}(\theta_*, \gamma_*) = 0, \tag{3.9}$$

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

As seen in Section 7, (3.9) is satisfied when $\boldsymbol{G}_\gamma$ is nested in $\boldsymbol{F}_\theta$.

4. THE VARIANCE STATISTIC

In the previous section, we show that whether the LR statistic is asymptotically distributed as a normal or a weighted sum of chi-squares depends on whether $f(\cdot\,|\,\cdot\,;\theta_*) = g(\cdot\,|\,\cdot\,;\gamma_*)$. Such a condition may hold when the conditional models $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ are nested or overlapping. It is therefore important to know if it is satisfied. Since θ_* and γ_* are unknown, we propose in this section a test of such a condition. The proposed test is based on the following property.

LEMMA 4.1: *Given Assumptions A1-(b), A2, A3, and A6,* $f(\cdot\,|\,\cdot\,;\theta_*) = g(\cdot\,|\,\cdot\,;\gamma_*)$ *if and only if* $\omega_*^2 = 0$.

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

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

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

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

Note that ω_*^2 is the variance of the limiting normal distribution of the LR statistic (Theorem 3.3-(ii)). Thus the variance statistic $\hat{\omega}_n^2$ plays two roles: first, to be a basis for a test of $f(\cdot\,|\,\cdot\,;\theta_*) = g(\cdot\,|\,\cdot\,;\gamma_*)$; 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 statistic is

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

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

LEMMA 4.2: *Given Assumptions A1–A3, and A6:*

$$\text{(4.4)} \qquad \text{(i)}\ \hat{\omega}_n^2 \xrightarrow{a.s.} \omega_*^2,$$

$$\text{(4.5)} \qquad \text{(ii)}\ \tilde{\omega}_n^2 \xrightarrow{a.s.} \omega_*^2 + \left[E^0 \left[\log \frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)} \right] \right]^2.$$

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_X^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_X^0-integrable functions independent 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:*

$$\text{(4.6)} \qquad n\hat{\omega}_n^2 = n\tilde{\omega}_n^2 + o_p(1) \xrightarrow{D} M_{p+q}(\cdot\,; \lambda_*^2)$$

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 parameters λ_*^2 are, as expected, all nonnegative.

As for the LR statistic, it is of interest to know when the limiting weighted sum of chi-square distribution of the variance statistics reduces 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 (3.8) hold.

COROLLARY 4.4 (Asymptotic Chi-Square Distribution of the Variance Statistics Given Information Matrix Equivalences and $\omega_*^2 = 0$): *Given Assumptions A1–A7, suppose that (3.8) holds. Then, under H_0^ω: $\omega_*^2 = 0$, the following statements 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, conditions (iii) or (iv) are satisfied if $\boldsymbol{G}_\gamma$ is nested in $\boldsymbol{F}_\theta$ or if $\boldsymbol{F}_\theta$ is nested in $\boldsymbol{G}_\gamma$. Conditions (iii) and (iv) can also be satisfied when the models are non-nested or overlapping. In particular, they are satisfied when the conditional models $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ are asymptotically orthogonal as defined by

Gourieroux, Monfort, and Trognon (1983), i.e., when:

$$B_{fg}(\theta_*, \gamma_*) = 0, \tag{4.7}$$

in which case the number of degrees of freedom 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 Sections 3 and 4, we shall obtain very simple tests for selecting among two competing models whether they are nested or misspecified. 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 the conditional model $\boldsymbol{F}_\theta$, this measure gives:

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

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

Given a pair 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 consider the following hypotheses and definitions:

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

meaning that $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ are *equivalent*, against

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

meaning that $\boldsymbol{F}_\theta$ is *better* than $\boldsymbol{G}_\gamma$, or

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

meaning that $\boldsymbol{F}_\theta$ is *worse* than $\boldsymbol{G}_\gamma$. These definitions have the desirable property that a correctly specified model must be at least as good as any other model.[3] Thus, if one rejects H_0 in favor of H_f, say, then $\boldsymbol{G}_\gamma$ must be misspecified.

[3] There are alternative definitions. For instance, one can use the mean-square error of prediction (see White and Olson (1979)). To take into account the parsimonious nature of a model, one may also adjust the above definitions by a correction factor $k(p, q)$ (see Vuong (1986, Theorem 5.4)). In this latter case, a correctly specified model is no longer necessarily best.

Another property is that the null hypothesis H_0 does not require that either of the competing models be correctly specified. As a matter of fact, from Lemma 6.2 below, both models must be misspecified under H_0 if $f(\cdot \mid \cdot\,; \theta_*) \neq g(\cdot \mid \cdot\,; \gamma_*)$.

The indicator $E^0[\log f(Y_t|Z_t; \theta_*)] - E^0[\log g(Y_t|Z_t; \gamma_*)]$ is, however, unknown. But we can consistently estimate this indicator by $(1/n)$ times the LR statistic under general conditions (Lemma 3.1). Thus the LR statistic is a natural statistic for discriminating between two models. Tests of H_0 against H_f or H_g will be called tests for model selection.

Since Cox's (1961, 1962) initial work, non-nested models have attracted a lot of interest (see, e.g., Mackinnon's (1983) recent survey and the special issue of the *Journal of Econometrics* edited by White (1983)). For a long time, non-nested hypotheses were defined as hypotheses that cannot be obtained from each other by a suitable limiting approximation (Cox (1961, 1962)). Noting that there were no satisfactory definitions, Pesaran (1987) recently proposed definitions of globally non-nested, partially non-nested, and nested hypotheses. It can be shown that Pesaran's definitions are equivalent to our Definitions 2, 3, and 4 below. Our definitions are more intuitive.

In this section, we consider the case where the models $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ are (strictly) non-nested. We give the following formal definition.

DEFINITION 2 (Strictly Non-Nested Models): Two conditional models $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ are *strictly non-nested* if and only if:

$$\boldsymbol{F}_\theta \cap \boldsymbol{G}_\gamma = \phi. \tag{5.5}$$

Since conditional distributions for Y_t given Z_t are defined on $\sigma_Y \times \boldsymbol{Z}$, and since some values of z may not be observed, condition (5.5) is to be understood as meaning that there is no conditional distribution for Y_t given Z_t which is equal to an element of $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ for H_Z^0-almost all z. A similar remark applies to Definitions 3 and 4 below. Condition (5.5) is satisfied when $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ are standard linear regression models with different distributional assumptions on the errors, say normally or logistic distributed. Alternatively, the competing regression models may have the same distributional assumption but different functional forms such as $Y_t = \theta_1 + \theta_2' Z_t + e_{ft}$ and $Y_t = \exp(\gamma_1 + \gamma_2' Z_t) + e_{gt}$ where $\theta_2 \neq 0$, $\gamma_2 \neq 0$, and Z_t is a nondegenerate real random vector.

Since the conditional models $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ do not have any conditional distribution in common, it must be the case that $f(\cdot \mid \cdot\,; \theta_*) \neq g(\cdot \mid \cdot\,; \gamma_*)$.[4] It follows that the second part of Theorem 3.3 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 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.

[4] For, if $f(y|z; \theta_*) = g(y|z; \gamma_*)$ holds for H_X^0-almost all (y, z), then from Assumption A1-(b) this must also hold for $(\nu_Y \times H_Z^0)$-almost all (y, z). Hence $F_{Y|Z}(\cdot|z; \theta_*) = G_{Y|Z}(\cdot|z; \gamma_*)$ for H_Z^0-almost all z, which implies a contradiction.

THEOREM 5.1 (Model Selection Tests for Strictly Non-Nested Models): *Given Assumptions A1–A6, if F_θ and G_γ are strictly non-nested, then*

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

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

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

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

Theorem 5.2 provides very simple directional tests 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 be based on the other statistic $n^{-1/2}LR_n(\hat{\theta}_n, \hat{\gamma}_n)/\tilde{\omega}_n$.

Both statistics are easy to compute. Each one is equal to the difference in the maximum log-likelihood values for the two models suitably normalized. The normalization $n^{1/2}\tilde{\omega}_n$ is obtained from the sum of squares of $m_t \equiv \log[f(Y_t|Z_t; \hat{\theta}_n)/g(Y_t|Z_t; \hat{\gamma}_n)]$, while the normalization $n^{1/2}\hat{\omega}_n$ is obtained from the sum of squared deviations of m_t from its sample mean which is equal to $(1/n)LR_n(\hat{\theta}_n, \hat{\gamma}_n)$. See (4.2)–(4.3). Alternatively, these statistics can be readily obtained from an additional linear regression. For instance, it can be shown that $n^{-1/2}LR_n(\hat{\theta}_n, \hat{\gamma}_n)/\hat{\omega}_n$ 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.[5]

The previous tests are based on the *unadjusted* LR statistic. There are, however, many equivalent statistics that can be used to form a model selection test. For instance, we may consider the following adjusted LR statistic:

$$\text{(5.9)} \quad \tilde{LR}_n(\hat{\theta}_n, \hat{\gamma}_n) \equiv LR_n(\hat{\theta}_n, \hat{\gamma}_n) - K_n(F_\theta, G_\gamma),$$

where $K_n(F_\theta, G_\gamma)$ is a correction factor depending on the characteristics of the competing models F_θ and G_γ such as their number of parameters. Suppose that:

$$\text{(5.10)} \quad n^{-1/2}K_n(F_\theta, G_\gamma) = o_p(1).$$

Examples of correction factors that satisfy (5.10) are $K_n(F_\theta, G_\gamma) = p - q$ and $K_n(F_\theta, G_\gamma) = (p/2)\log n - (q/2)\log n$, which correspond to Akaike (1973) and Schwarz (1978) information criteria. It is clear that $n^{-1/2}\tilde{LR}_n(\hat{\theta}_n, \hat{\gamma}_n)/\hat{\omega}_n$ has the same asymptotic properties as $n^{-1/2}LR_n(\hat{\theta}_n, \hat{\gamma}_n)/\hat{\omega}_n$. Hence, we can use the adjusted log-likelihood ratio $\tilde{LR}_n(\hat{\theta}_n, \hat{\gamma}_n)$ as a basis for a model selection test. In

[5] I owe this point to Hal White.

terms of the unadjusted LR statistic, we would accept H_0 whenever

$$-c + n^{-1/2}K_n(F_\theta, G_\gamma)/\hat{\omega}_n \leqslant n^{-1/2}LR_n(\hat{\theta}_n, \hat{\gamma}_n)/\hat{\omega}_n \leqslant c + n^{-1/2}K_n(F_\theta, G_\gamma)/\hat{\omega}_n$$

where c is obtained from the standard normal distribution. Thus 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 $n^{-1/2}L\tilde{R}_n(\hat{\theta}_n, \hat{\gamma}_n)/\hat{\omega}_n$ is approximated under H_0 by the asymptotic $N(0,1)$ distribution. In the next sections on overlapping models and nested models, we shall not discuss possible adjustments to the LR statistic. Similar results can clearly be established.

We now contrast our approach to those initiated by Akaike (1973, 1974) on model selection and Cox (1961, 1962) on non-nested hypothesis testing. First, the difference between Akaike's and our approach is that ours 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. As in the classical testing situation, our distributional results are used to indicate the strength of the evidence in favor of either model whether it is based on the adjusted or unadjusted LR statistic. As a consequence we do not have to choose a "best" model if the competing models are statistically equivalent.

Second, the difference between Cox's and our approach lies in the null hypotheses under test. In Cox's approach, the implicit null hypothesis when testing F_θ using the evidence providing by G_γ, say, is:

$$(5.11) \quad H_0^f\colon E^0\left[\log \frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right] = \int_Z\int_Y \log \frac{f(y|z;\theta_*)}{g(y|z;\gamma_*)} f(y|z;\theta_*)\nu_Y(dy)H_Z^0(dz)$$

(see Aguirre-Torres and Gallant (1983), White (1982b)). Hence H_0^f and H_0 are in general different. As a matter of fact, these null hypotheses are identical if and only if $f(\cdot\,|\,\cdot\,;\theta_*) = g(\cdot\,|\,;\gamma_*)$, which cannot hold when the models are strictly non-nested. Moreover, it is well-known that H_0^f holds if F_θ is correctly specified. On the other hand, as noticed earlier, when the competing models are strictly non-nested, both models must be misspecified under our null hypothesis H_0.

6. OVERLAPPING MODELS

In this section, we consider 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. We first give a formal definition of overlapping models.

DEFINITION 3 (Overlapping Models): Two conditional models $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ are *overlapping* if and only if:

(6.1) (i) $\boldsymbol{F}_\theta \cap \boldsymbol{G}_\gamma \neq \phi$,

(6.2) (ii) $\boldsymbol{F}_\theta \not\subset \boldsymbol{G}_\gamma$ and $\boldsymbol{G}_\gamma \subset \boldsymbol{F}_\theta$.

Condition (i) says that $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ have some common conditional distributions for Y_t given Z_t for H_Z^0-almost all z, while condition (ii) says that neither model is 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. Since $\boldsymbol{F}_\theta \cap \boldsymbol{G}_\gamma \neq \phi$, then one may have $f(\cdot \mid \cdot\,; \theta_*) = g(\cdot \mid \cdot\,; \gamma_*)$. From Theorem 3.3, it follows that under the null hypothesis H_0:

(6.3) (i) if $f(\cdot \mid \cdot\,; \theta_*) = g(\cdot \mid \cdot\,; \gamma_*), 2LR_n(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{D} M_{p+q}(\cdot, \lambda_*)$,

(6.4) (ii) if $f(\cdot \mid \cdot\,; \theta_*) \neq g(\cdot \mid \cdot\,; \gamma_*), n^{-1/2}LR_n(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{D} N(0, \omega_*^2)$.

For instance, in the normal linear regression case with some common explanatory variables, (i) occurs if and only if the pseudo-true parameters associated with the variables specific to each regression are simultaneously null (see Lien and Vuong (1987)). A stronger condition is that $H^0_{Y|Z}(\cdot \mid \cdot)$ is common to the two competing linear models, or equivalently that both linear models are correctly specified. Since one does not know a priori if $f(\cdot \mid \cdot\,; \theta_*) = g(\cdot \mid \cdot\,; \gamma_*)$ holds, i.e., if the distributions in $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ that are closest to $H^0_{Y|Z}(\cdot \mid \cdot)$ are observationally identical, 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 $f(\cdot \mid \cdot\,; \theta_*) = g(\cdot \mid \cdot\,; \gamma_*)$ 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 $f(\cdot \mid \cdot\,; \theta_*) = g(\cdot \mid \cdot\,; \gamma_*)$ 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 $\tilde{\omega}_n^2$.[6] Such a test is called the variance test. Once it is known whether $\omega_*^2 = 0$, 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 carry out a test of H_0 against H_f or H_g when $\omega_*^2 = 0$. Indeed H_0^ω is clearly included in H_0 so that $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$

[6]An alternative to the variance test is to characterize and test the conditions that θ_* and γ_* must satisfy for $f(\cdot \mid \cdot\,; \theta_*)$ to be equal to $g(\cdot \mid \cdot\,; \gamma_*)$. See Lien and Vuong (1987) for an illustration. In general, tests of these conditions are easier to perform and can be done using $\hat{\theta}_n$ and $\hat{\gamma}_n$.

must necessarily be equivalent when $\omega_*^2 = 0$. On the other hand, when $\omega_*^2 \neq 0$, 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. But, when $\omega_*^2 \neq 0$, (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 $n^{-1/2}LR_n(\hat{\theta}_n, \hat{\gamma}_n)/\tilde{\omega}_n$ discussed in Section 5 applies.

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 is not rejected, conclude that F_θ and G_γ cannot be discriminated given the data. If H_0^ω is rejected, (ii) test H_0 against H_f or H_g using the normal model selection test based on $n^{-1/2}LR_n(\hat{\theta}_n, \hat{\gamma}_n)/\hat{\omega}_n$ or $n^{-1/2}LR_n(\hat{\theta}_n, \hat{\gamma}_n)/\tilde{\omega}_n$ 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 a 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 and the normal LR-test. To see this, note that H_0 is a composite of H_0^ω and $H_0 - H_0^\omega$. Let $A \equiv \{n\hat{\omega}_n^2 > c_1\}$ and $B \equiv \{|n^{-1/2}LR(\hat{\theta}_n, \hat{\gamma}_n)/\hat{\omega}_n| > c_2\}$. Then $\Pr[\text{reject } H_0|H_0] = \Pr[A \cap B|H_0] \leqslant \max\{\Pr(A \cap B|H_0^\omega), \Pr(A \cap B|H_0 - H_0^\omega)\} \leqslant \max\{\Pr(A|H_0^\omega), \Pr(B|H_0 - H_0^\omega)\}$. But from Theorems 5.1 and 6.1 below, $\Pr(A|H_0^\omega) \to \alpha_1$ and $\Pr(B|H_0 - H_0^\omega) \to \alpha_2$. Thus if $\alpha_1 = \alpha_2 = .10$, the significance level of the 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 (3.6). For instance, $\hat{W}_n$ is obtained by replacing in (3.6) the matrix $B_{fg}(\theta_*, \gamma_*)$, say, by its sample analog $B_{fgn}(\hat{\theta}_n, \hat{\gamma}_n)$ defined in (2.7). Let $\hat{\lambda}_n^2$ be the vector of squares of $\hat{\lambda}_n$.

THEOREM 6.1 (Variance Tests for Discrimination): *Given Assumptions A1–A7,*
(i) *under* H_0^ω, *for any* $x \geqslant 0$, $\Pr(n\hat{\omega}_n^2 \leqslant x) - M_{p+q}(x; \hat{\lambda}_n^2) \xrightarrow{a.s.} 0$,
(ii) *under* H_A^ω, $n\hat{\omega}_n^2 \xrightarrow{a.s.} +\infty$,
(iii) *properties* (*i*) *and* (*ii*) *hold for* $n\tilde{\omega}_n^2$.

The variance test consists in choosing a critical value x so that $M_{p+q}(x; \hat{\lambda}_n^2) = 1 - \alpha\%$ for some significance level α, and in rejecting H_0^ω if $n\hat{\omega}_n^2 > x$.[7] Part (i) ensures that the asymptotic size is α, while (ii) says that the test is consistent. Similar conclusions apply to the test based on $n\tilde{\omega}_n^2$. Computation of the statistic $n\hat{\omega}_n^2$ and $n\tilde{\omega}_n^2$ is straightforward. The test also requires the computation of the eigenvalues $\hat{\lambda}_n$. The eigenvalues need not, however, be computed when rank $B_{fg}(\theta_*, \gamma_*)$ is known and condition (iii) or (iv) of Corollary 4.4. holds. (Orthogonal models fullfill such requirements.) In this case, both $n\hat{\omega}_n^2$ and $n\tilde{\omega}_n^2$ converge, under H_0^ω, to a chi-square distribution with degrees of freedom equal to $p + q - 2\,\text{rank}\, B_{fg}(\theta_*, \gamma_*)$.

[7]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 $M_m(x; \lambda)$. Dubin and Rivers (1986) also have an efficient and flexible subroutine for computing $M_m(x; \lambda)$.

As pointed our earlier, the difficulty in selecting among overlapping models arises from the fact that $f(\cdot \mid \cdot\,; \theta_*)$ may or may not be equal to $g(\cdot \mid \cdot\,; \gamma_*)$ under the null hypothesis of interest H_0 so that the form of the asymptotic null distribution of the LR statistic is a priori unknown. This is not the case if one knows that at least one of the two overlapping models is correctly specified, a frequent assumption in the model selection literature. We say that the conditional model $\boldsymbol{F}_\theta$, for instance, is correctly specified if $H^0_{Y|Z}(\cdot \mid \cdot) \in \boldsymbol{F}_\theta$, i.e., if there exists a θ_o in Θ such that $H^0_{Y|Z}(\cdot \mid z) = F_{Y|Z}(\cdot \mid z; \theta_o)$ for H^0_Z-almost all z.

LEMMA 6.2: *Given Assumptions A1-(b), A2 and A3, suppose that* $\boldsymbol{F}_\theta$ *and* $\boldsymbol{G}_\gamma$ *are overlapping and at least one model is correctly specified. Then the following statements are equivalent*:

(i) $H^0_{Y|Z}(\cdot \mid \cdot) \in \boldsymbol{F}_\theta \cap \boldsymbol{G}_\gamma$,

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

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

From (i) and (iii) it follows that, when at least one model is correctly specified, then the models $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ are equivalent if and only if the other model is correctly specified. From (ii) and (iii) we have that the models $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ are equivalent if and only if $f(\cdot \mid \cdot\,; \theta_*) = g(\cdot \mid \cdot\,; \gamma_*)$. The importance of this second equivalence is that under H_0, we now always have $f(\cdot \mid \cdot\,; \theta_*) = g(\cdot \mid \cdot\,; \gamma_*)$ so that the asymptotic distribution of the LR statistic is given by the weighted sum of chi-squares obtained in Theorem 3.3-(i). Thus in this case we can bypass the above sequential procedure, and directly construct a model selection test based on the LR statistic.

THEOREM 6.3 (Model Selection Test for Overlapping Models): *Given Assumptions A1–A5, if* $\boldsymbol{F}_\theta$ *and* $\boldsymbol{G}_\gamma$ *are overlapping and at least one model is correctly specified, then*:

(i) *under* H_0, *for any* $x \geqslant 0$, $\Pr(2LR_n(\hat{\theta}_n, \hat{\gamma}_n) \leqslant x) - M_{p+q}(x; \hat{\lambda}_n) \xrightarrow{a.s.} 0$,

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

(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 distribution. 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. Since at least one model is assumed to be correctly specified, then rejection of H_0 in favor of H_f, say, implies that $\boldsymbol{F}_\theta$ is correctly specified and $\boldsymbol{G}_\gamma$ is misspecified. The test requires consistent estimators $\hat{\lambda}_n$ of λ_*. If the competing models are asymptotically orthogonal, it can readily be shown from (3.6) that λ_* is equal to a vector of p ones and q minus ones so that the limiting distribution reduces to that 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.

A formal definition of nested models is:

DEFINITION 4 (Nested Models): The conditional model $\boldsymbol{G}_\gamma$ is *nested* in $\boldsymbol{F}_\theta$ if and only if:

$$\boldsymbol{G}_\gamma \subset \boldsymbol{F}_\theta. \tag{7.1}$$

As before, condition (7.1) means that any conditional distribution in $\boldsymbol{G}_\gamma$ is equal to a conditional distribution in $\boldsymbol{F}_\theta$ for H_Z^0-almost all z. We make the following regularity assumption on the parameterizations θ and γ.

ASSUMPTION A8: *There exists a C^2-function $\phi(\cdot)$ from Γ to Θ such that for any γ in Γ:*

$$g(\cdot \mid \cdot\,; \gamma) = f(\cdot \mid \cdot\,; \phi(\gamma)) \quad \textit{for } \left(\nu_Y \times H_Z^0\right)\textit{—almost } \forall\, (y, z). \tag{7.2}$$

Assumption A8 states that any conditional density $g(\cdot \mid \cdot\,; \gamma)$ is also a conditional density $f(\cdot \mid \cdot\,; \theta)$ for some θ in Θ. Since $\phi(\Gamma)$ is included in Θ, (7.1) holds so that $\boldsymbol{G}_\gamma$ is nested in $\boldsymbol{F}_\theta$.

The pseudo-true parameter θ_* is not necessarily equal to $\phi(\gamma_*)$ since θ_* may not belong to $\phi(\Gamma)$. The next result relates the condition $\theta_* \in \phi(\Gamma)$ to the condition that $\boldsymbol{F}_\theta$ and $\boldsymbol{G}_\gamma$ are equivalent, and to the condition that $f(\cdot \mid \cdot\,; \theta_*) = g(\cdot \mid \cdot\,; \gamma_*)$ for H_X^0-almost all (y, z).

LEMMA 7.1: *Given Assumptions A1-(b), A2–A3, and A8, the following statements are equivalent:*

(i) $\theta_* = \phi(\gamma_*)$,
(ii) $\theta_* \in \phi(\Gamma)$,
(iii) $E^0[\log f(Y_t \mid Z_t; \theta_*)] = E^0[\log g(Y_t \mid Z_t; \gamma_*)]$,
(iv) $f(\cdot \mid \cdot\,; \theta_*) = g(\cdot \mid \cdot\,; \gamma_*)$.

Lemma 7.1 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. 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^θ: $\theta_* \notin \phi(\Gamma)$ and $H_f \cup H_g$. As a matter of fact, the alternative to the null hypothesis H_0 is H_f since H_g can never occur

because G_γ can never be better than F_θ. Hence, Lemma 7.1 says that testing whether or not θ_* satisfies some restrictions is equivalent to testing whether the smaller model is equivalent to or worse than the larger model.

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 7.1, we always have $f(\cdot\,|\,\cdot\,;\theta_*) = g(\cdot\,|\,\cdot\,;\gamma_*)$ under the null hypothesis H_0. Hence, there is no ambiguity as to the asymptotic distribution of the LR statistic which is the weighted sum of chi-squares obtained in Theorem 3.3-(i). 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:

$$(7.3) \qquad LR_n \equiv \sum_{t=1}^{n} \log \frac{f(Y_t|Z_t;\hat{\theta}_n)}{f(Y_t|Z_t;\tilde{\theta}_n)} = LR_n(\hat{\theta}_n, \hat{\gamma}_n),$$

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

The next result is similar to Kent's (1982) Theorem 3.1, and gives the properties of the model selection or nested hypothesis test based on the LR statistic. It simplifies the computation of the nonzero eigenvalues of W by replacing W by a matrix $\underline{W}$ of lower dimension. Let:

$$(7.4) \qquad \underline{W} = B_f(\theta_*)\left[\frac{\partial\phi(\gamma_*)}{\partial\gamma'} A_g^{-1}(\gamma_*) \frac{\partial\phi'(\gamma_*)}{\partial\gamma} - A_f^{-1}(\theta_*)\right],$$

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

THEOREM 7.2 (LR Tests for Nested Models): *Given Assumptions A1–A5 and A8, the eigenvalues* $\hat{\underline{\lambda}}_n$ *are almost surely all real nonnegative and*:

(i) *under* H_0^θ, *for any* $x \geqslant 0$, $\Pr(2LR_n \leqslant x) - M_p(x; \hat{\underline{\lambda}}_n) \xrightarrow{a.s.} 0$,

(ii) *under* H_A^θ, $2LR_n \xrightarrow{a.s.} +\infty$.

The test is one sided. It is carried out by choosing a critical value from $M_p(\cdot\,;\hat{\underline{\lambda}}_n)$ 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 equivalence holds for the larger model, one has the following corollary.

COROLLARY 7.3 (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, $2LR_n \xrightarrow{D} \chi^2_{p-q}$,

(ii) *under* H_A^θ, $2LR_n \xrightarrow{a.s.} +\infty$.

The well-known Wilks (1938) result follows since the information matrix equivalence holds if the larger model is correctly specified.

Using the equivalence between H_0^θ and H_0, we have motivated the LR statistic as a basis for a test of H_0^θ against H_A^θ under general conditions. From Lemmas 7.1 and 4.1, we also have the equivalence between H_0^θ and H_0^ω: $\omega_*^2 = 0$. This suggests testing the parametric hypothesis H_0^θ against H_A^θ by testing H_0^ω against H_A^ω. Thus, we have a new test for nested hypotheses based on the variance statistics $\hat{\omega}_n^2$ and $\tilde{\omega}_n^2$. Let $\hat{\underline{\lambda}}_n^2$ be the squares of the eigenvalues $\hat{\underline{\lambda}}_n$.

THEOREM 7.4 (Variance Tests for Nested Models): *Given Assumptions A1–A8:*
(i) *under* H_0^θ, *for any* $x \geqslant 0, \Pr(n\hat{\omega}_n^2 \leqslant x) - M_p(x; \hat{\underline{\lambda}}_n^2) \xrightarrow{a.s.} 0$,
(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 $M_p(\cdot; \hat{\underline{\lambda}}_n^2)$ 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:

$$n\hat{\omega}_n^2 = \sum_{t=1}^{n} \left[\log \frac{f(Y_t|Z_t; \hat{\theta}_n)}{f(Y_t|Z_t; \tilde{\theta}_n)} \right]^2 - \frac{1}{n} LR_n^2, \tag{7.5}$$

$$n\tilde{\omega}_n^2 = \sum_{t=1}^{n} \left[\log \frac{f(Y_t|Z_t; \hat{\theta}_n)}{f(Y_t|Z_t; \tilde{\theta}_n)} \right]^2. \tag{7.6}$$

Thus $n\hat{\omega}_n^2$ is the sum of squared residuals in a linear regression of

$$m_t \equiv \log \left[f(Y_t|Z_t; \hat{\theta}_n) / f(Y_t|Z_t; \tilde{\theta}_n) \right]$$

on the constant term. The variance tests are not asymptotically equivalent to the LR tests, and require more assumptions than the LR test. In normal linear regressions, these additional assumptions bear on the fourth moments of the residuals. Thus it is expected that the variance statistics would be less stable than the LR statistic.[8]

If the larger model is correctly specified, then the limiting distribution reduces to the central chi-square distribution with $p - q$ degrees of freedom, as other classical statistics.

COROLLARY 7.5 (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, $n\hat{\omega}_n^2 \xrightarrow{D} \chi_{p-q}^2$,
(ii) *under* H_A^θ, $n\hat{\omega}_n^2 \xrightarrow{a.s.} +\infty$,
(iii) *properties* (*i*) *and* (*ii*) *hold for* $n\tilde{\omega}_n^2$.

[8] The variance tests are neither asymptotically equivalent under H_0^θ to the robust Wald and LM tests proposed by White (1982a). The asymptotic power properties of the variance tests in the misspecified case are left for future research.

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. The 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 KLIC, all our model selection tests are based on the LR statistic. As a prerequisite, we have fully characterized the asymptotic distribution of the LR statistic under the most general conditions.

Much work remains to be done. First, one could relax Assumption A1-(a) so as to extend our results to time-series models. Second, we have mentioned that our LR-based tests for model selection could be adjusted for the number of parameters. A theoretical and Monte Carlo study would shed some light on the most adequate adjustment to the LR statistic in small samples for some particular cases. Third, a thorough comparison between our model selection tests and the available Cox-type tests as considered by Davidson and McKinnon (1981), Pesaran (1974), and Pesaran and Deaton (1978), among others, would be useful. In the same line, it would be useful to compare our approach to the comprehensive approach advocated by Atkinson (1969, 1970), which requires nesting the competing models in a larger model. 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 (1986). Fifth, the above model selection tests have been obtained under the assumption that there are only two competing models. It is important to generalize our procedures to the case where there are many competing models.

Department of Economics, University of Southern California, Los Angeles, CA 90089, U.S.A.

Manuscript received March, 1986; revision received January, 1988.

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_*. 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. The following lemma is useful.

LEMMA A: *Given Assumptions A1–A5:*

$$(A.1)\qquad n^{1/2}\begin{bmatrix}\hat{\theta}_n-\theta_*\\ \hat{\gamma}_n-\gamma_*\end{bmatrix}\xrightarrow{D} N(0,\Sigma),\quad \textit{where}\quad \Sigma=\begin{bmatrix}A_f^{-1}B_fA_f^{-1}; & A_f^{-1}B_{fg}A_g^{-1}\\ A_g^{-1}B_{gf}A_f^{-1}; & A_g^{-1}B_gA_g^{-1}\end{bmatrix}.$$

Moreover, the asymptotic covariance matrix Σ can be consistently estimated by $\hat{\Sigma}_n$ which is defined as in (A.1) where A_s, B_s, B_{fg}, $s=f,g$, are replaced by their sample analogs evaluated at the ML estimators $\hat{\theta}_n$ and $\hat{\gamma}_n$.

PROOF OF LEMMA A: Given Assumptions A1–A5, we obtain using the Taylor expansions of the normal equations:

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

$$(A.3)\qquad 0=n^{-1/2}\frac{\partial L_n^g(\gamma_*)}{\partial\gamma}+A_g\cdot n^{1/2}(\hat{\gamma}_n-\gamma_*)+o_p(1).$$

On the other hand from the multivariate Central Limit Theorem (see, e.g., Rao (1973)):

$$(A.4)\qquad n^{-1/2}\begin{bmatrix}\partial L_n^f(\theta_*)/\partial\theta\\ \partial L_n^g(\gamma_*)/\partial\gamma\end{bmatrix}\xrightarrow{D} N\left(0,\begin{bmatrix}B_f; & B_{fg}\\ B_{gf}; & B_g\end{bmatrix}\right).$$

The desired result follows since A_f and A_g are nonsingular (White (1982a, Theorem 3.1)).

PROOF OF LEMMA 3.1: Obvious from, e.g., Vuong (1983, Theorem 1).

PROOF OF LEMMA 3.2: From Moore (1978, Theorem 1), we know that $Y'QY\sim M_m(\cdot;\lambda)$ where λ are the eigenvalues of $\Omega^{1/2}Q\Omega^{1/2}$ and $\Omega^{1/2}$ is the (unique) square root of Ω. Using Theorem 1.3.20 in Horn and Johnson (1985) it follows that the eigenvalues of $\Omega^{1/2}Q\Omega^{1/2}$ are the eigenvalues of $Q\Omega$.

PROOF OF THEOREM 3.3: From a Taylor expansion of $L_n^f(\theta_*)$ around $\hat{\theta}_n$, we obtain:

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

Similarly, we have:

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

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

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

To prove (i), we note that $LR_n(\theta_*,\gamma_*)=0$ if $f(\cdot\,|\,\cdot;\theta_*)=g(\cdot\,|\,\cdot;\gamma_*)$. Part (i) follows from Lemma A and Lemma 3.2 by considering the quadratic form associated with the block-diagonal matrix:

$$(A.8)\qquad Q=\begin{bmatrix}-A_f & 0\\ 0 & A_g\end{bmatrix}.$$

Then, one can check that $Q\Sigma$ is equal to W as given in (3.6). To prove (ii), we note that $n^{1/2}(\hat{\theta}_n-\theta_*)$ and $n^{1/2}(\hat{\gamma}_n-\gamma_*)$ are $O_p(1)$. Thus, from (A.7) we obtain:

$$(A.9)\qquad n^{-1/2}LR_n(\hat{\theta}_n,\hat{\gamma}_n)-n^{1/2}E^0\left[\log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right]$$
$$=n^{1/2}\left[\frac{1}{n}LR_n(\theta_*,\gamma_*)-E^0\left[\log\frac{f(Y_t|Z_t;\theta_*)}{g(Y_t|Z_t;\gamma_*)}\right]\right]+o_p(1).$$

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 (3.4). This variance is finite given Assumption A6 and the Cauchy-Schwartz inequality. Part (ii) follows.

PROOF OF COROLLARY 3.4: From the proof of Theorem 3.3-(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 $N(0, \Sigma)$ (Lemma A). 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 $\Sigma Q \Sigma Q \Sigma = \Sigma Q \Sigma$, where Q is given in (A.8), in which case the number of degrees of freedom is $\operatorname{tr} Q\Sigma$. But $\Sigma = A^{-1}BA^{-1}$, where

$$\text{(A.10)} \qquad B = \begin{bmatrix} B_f & B_{fg} \\ B_{gf} & B_g \end{bmatrix}; \quad A = \begin{bmatrix} A_f & 0 \\ 0 & A_g \end{bmatrix}.$$

Noticing that $A^{-1}QA^{-1} = Q^{-1}$, the necessary and sufficient condition $\Sigma Q \Sigma Q \Sigma = \Sigma Q \Sigma$ becomes:

$$\text{(A.11)} \qquad BQ^{-1}BQ^{-1}B = BQ^{-1}B.$$

Using (3.8), we obtain that (A.11) is equivalent to:

$$\text{(A.12)} \qquad \begin{bmatrix} B_f - B_{fg}B_g^{-1}B_{gf}; & B_{fg}B_g^{-1}\left(B_g - B_{gf}B_f^{-1}B_{fg}\right) \\ \left(B_g - B_{gf}B_f^{-1}B_{fg}\right)B_g^{-1}B_{gf}; & B_g - B_{gf}B_f^{-1}B_{fg} \end{bmatrix}$$
$$= \begin{bmatrix} B_f - B_{fg}B_g^{-1}B_{gf}; & 0 \\ 0; & -B_g + B_{gf}B_f^{-1}B_{fg} \end{bmatrix},$$

which is equivalent to (3.9). Then, the number of degrees of freedom is:

$$\text{(A.13)} \qquad \operatorname{tr} Q\Sigma = \operatorname{tr}\begin{bmatrix} I_p; & B_{fg}B_g^{-1} \\ -B_{gf}B_f^{-1}; & -I_q \end{bmatrix} = p - q.$$

PROOF OF LEMMA 4.1: From (3.4) it follows that $\omega_*^2 = 0$ if and only if $f(\cdot \mid \cdot\,; \theta_*) = Kg(\cdot \mid \cdot\,; \gamma_*)$ for some constant K, H_X^0-almost surely. It remains to show that $K = 1$. From Assumption A1-(b), it follows that $f(y \mid z; \theta_*) = Kg(y \mid z; \gamma_*)$ for $(\nu_Y \times H_Z^0)$-almost all (y, z). Integrating this equality with respect to $(\nu_Y \times H_Z^0)$ gives $1 = K$.

PROOF OF LEMMA 4.2: Given Assumptions A1–A3, and A6, it follows from the Cauchy-Schwartz inequality and Jennrich's uniform Strong Law of Large Numbers (1969, Theorem 2) that

$$\text{(A.14)} \qquad \frac{1}{n}\sum_{t=1}^{n}\left[\log\frac{f(Y_t \mid Z_t; \theta)}{g(Y_t \mid Z_t; \gamma)}\right]^2 \xrightarrow{a.s.} E^0\left[\log\frac{f(Y_t \mid Z_t; \theta)}{g(Y_t \mid Z_t; \gamma)}\right]^2,$$

uniformly in (θ, γ) on $\Theta \times \Gamma$. The result follows from Lemma 3.1 and the strong consistency of $\hat{\theta}_n$ and $\hat{\gamma}_n$ to θ_* and γ_*.

PROOF OF THEOREM 4.3: Since $\omega_*^2 = 0$ is equivalent to $f(\cdot \mid \cdot\,; \theta_*) = g(\cdot \mid \cdot\,; \gamma_*)$ (Lemma 4.1), it follows from Theorem 3.3-(i) that $LR_n(\hat{\theta}_n, \hat{\gamma}_n) = O_p(1)$. Thus, from (4.3), the equality in (4.6) follows. 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:

$$\text{(A.15)} \qquad \tilde{\omega}_n^2 = \frac{1}{n}\sum_{t=1}^{n}\left[\log\frac{f_t(\theta_*)}{g_t(\gamma_*)}\right]^2 + 2\left[\frac{1}{n}\sum_{t=1}^{n}\left[\log\frac{f_t(\theta_*)}{g_t(\gamma_*)}\right]\frac{\partial \log f_t(\theta_*)}{\partial \theta'}\right]\left(\hat{\theta}_n - \theta_*\right)$$
$$- 2\left[\frac{1}{n}\sum_{t=1}^{n}\left[\log\frac{f_t(\theta_*)}{g_t(\gamma_*)}\right]\frac{\partial \log g_t(\gamma_*)}{\partial \gamma'}\right](\hat{\gamma}_n - \gamma_*)$$
$$+ \left(\hat{\theta}_n' - \theta_*', \hat{\gamma}_n' - \gamma_*'\right)\bar{V}_n\left(\hat{\theta}_n' - \theta_*', \hat{\gamma}_n' - \gamma_*'\right)'$$

where 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

$$\text{(A.16)}\qquad \bar{V}_n = \begin{bmatrix} \bar{V}_{\theta\theta n}; & \bar{V}_{\theta\gamma n} \\ \bar{V}_{\gamma\theta n}; & \bar{V}_{\gamma\gamma n} \end{bmatrix},$$

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

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

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

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.15) are null. Moreover, given Assumptions 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 $O_p(1)$, we obtain from (A.15):

$$\text{(A.17)}\qquad n\tilde{\omega}_n^2 = n\left(\hat{\theta}'_n - \theta'_*, \hat{\gamma}_n - \gamma'_*\right) V \left(\hat{\theta}'_n - \theta'_*, \hat{\gamma}'_n - \gamma'_*\right) + o_p(1)$$

where

$$\text{(A.18)}\qquad V = \begin{bmatrix} B_f & -B_{fg} \\ -B_{gf} & B_g \end{bmatrix}.$$

From Lemma A and Lemma 3.2, it remains to show that the eigenvalues of $V\Sigma$ are equal to the squares of the eigenvalues of $W = Q\Sigma$ where Q is defined in (A.8). It is easy to check that $V = Q\Sigma Q$. Thus $V\Sigma = (Q\Sigma)^2$. This completes the proof.

PROOF OF COROLLARY 4.4: From Lemma A, (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 $\operatorname{tr} V\Sigma$. Using (3.8) we have:

$$\text{(A.19)}\qquad V\Sigma = \begin{bmatrix} I_p - B_{fg}B_g^{-1}B_{gf}B_f^{-1}; & 0 \\ 0; & I_q - B_{gf}B_f^{-1}B_{fg}B_g^{-1} \end{bmatrix},$$

$$\text{(A.20)}\qquad \Sigma V \Sigma = \begin{bmatrix} B_f^{-1}\left(I_p - B_{fg}B_g^{-1}B_{gf}B_f^{-1}\right); & B_f^{-1}B_{fg}B_g^{-1}\left(I_q - B_{gf}B_f^{-1}B_{fg}B_g^{-1}\right) \\ B_g^{-1}B_{gf}B_f^{-1}\left(I_p - B_{fg}B_g^{-1}B_{gf}B_f^{-1}\right); & B_g^{-1}\left(I_q - B_{gf}B_f^{-1}B_{fg}B_g^{-1}\right) \end{bmatrix},$$

$$\text{(A.21)}\qquad \Sigma V \Sigma V \Sigma = \begin{bmatrix} B_f^{-1}\left(I_p - B_{fg}B_g^{-1}B_{gf}B_f^{-1}\right)^2; & B_f^{-1}B_{fg}B_g^{-1}\left(I_q - B_{gf}B_f^{-1}B_{fg}B_g^{-1}\right)^2 \\ B_g^{-1}B_{gf}B_f^{-1}\left(I_p - B_{fg}B_g^{-1}B_{gf}B_f^{-1}\right)^2; & B_g^{-1}\left(I_q - B_{gf}B_f^{-1}B_{fg}B_g^{-1}\right)^2 \end{bmatrix}.$$

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 *both* idempotent, i.e., 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 *both* 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, $\operatorname{rank}(B_{fg}B_g^{-1})(B_{gf}B_f^{-1}) = \operatorname{rank} B_{fg}B_g^{-1}B_{gf} = \operatorname{rank} B_{fg} = \operatorname{rank} B_{fg}B_g^{-1}$. Thus, from Rao and Mitra (1971, Lemma 2.2.7), 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 equivalences between (i), (ii), (iii), and (iv). Finally, from (A.19):

$$\text{(A.22)} \qquad \operatorname{tr} V\Sigma = p + q - \operatorname{tr}\left(B_{fg}B_g^{-1}B_{gf}B_f^{-1}\right) - \operatorname{tr}\left(B_{gf}B_f^{-1}B_{fg}B_g^{-1}\right)$$

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

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 $\operatorname{tr}(B_{fg}B_g^{-1}B_{gf}B_f^{-1}) = \operatorname{rank}(B_{fg}B_g^{-1}B_{gf}B_f^{-1}) = \operatorname{rank} B_{gf}$. This completes the proof.

PROOF OF THEOREM 5.1: Straightforward from Theorem 3.3-(ii), and Lemma 4.2 since $f(\cdot\,|\,\cdot\,;\theta_*) \neq g(\cdot\,|\,\cdot\,;\gamma_*)$ (footnote 3) and $\omega_*^2 > 0$ (Lemma 4.1).

PROOF OF THEOREM 6.1: Since $\hat{W}_n$ converges almost surely to W and since the eigenvalues of a matrix are continuous in the elements of the matrix (see, e.g., Horn and Johnson (1985, p. 540)), then the eigenvalues $\hat{\lambda}_n$ converge almost surely to λ_*. Part (i) follows from Theorem 4.3, since $M_{p+q}(x;\lambda)$ is continuous in λ. Part (ii) follows from Lemma 4.2-(i). Part (iii) is proved similarly.

PROOF OF LEMMA 6.2: We shall prove that (ii) $\Rightarrow$ (i) $\Rightarrow$ (iii) $\Rightarrow$ (ii). We shall also freely switch between the measures $(\nu_Y \times H_Z^0)$ and H_X^0 because they are equivalent by Assumption A1-(b). Without loss of generality, we assume that $H_{Y|Z}^0(\cdot\,|\,\cdot) \in \boldsymbol{F}_\theta$, which is equivalent to $h^0(\cdot\,|\,\cdot) = f(\cdot\,|\,\cdot\,;\theta_o)$ for some θ_o in Θ. It follows from Assumption A3-(b) and Jensen's inequality that $\theta_* = \theta_o$. Thus $h^0(\cdot\,|\,\cdot) = f(\cdot\,|\,\cdot\,;\theta_*)$.

(ii) $\Rightarrow$ (i): Since $h^0(\cdot\,|\,\cdot) = f(\cdot\,|\,\cdot\,;\theta_*)$, then $h^0(\cdot\,|\,\cdot) = g(\cdot\,|\,\cdot\,;\gamma_*)$, so that $H_{Y|Z}^0(\cdot\,|\,\cdot) \in \boldsymbol{G}_\gamma$.

(i) $\Rightarrow$ (iii): Since $H_{Y|Z}^0(\cdot\,|\,\cdot) \in \boldsymbol{G}_\gamma$, then $h^0(\cdot\,|\,\cdot) = g(\cdot\,|\,\cdot\,;\gamma_*)$ as above. Since $h^0(\cdot\,|\,\cdot) = f(\cdot\,|\,\cdot\,;\theta_*)$, then $f(\cdot\,|\,\cdot\,;\theta_*) = g(\cdot\,|\,\cdot\,;\gamma_*)$ for H_X^0-almost all (y, z), which implies (iii).

(iii) $\Rightarrow$ (ii): Since $h^0(\cdot\,|\,\cdot) = f(\cdot\,|\,\cdot\,;\theta_*)$ for $(\nu_Y \times H_Z^0)$ almost all (y, z), (iii) implies:

$$\int_Z \left\{ \int_Y \log \frac{f(y|z;\theta_*)}{g(y|z;\gamma_*)} f(y|z;\theta_*)\nu_Y(dy) \right\} H_Z^0(dz) = 0.$$

Then (ii) follows from Jensen's inequality and $H_X^0 = F_{Y|Z}(\cdot\,|\,\cdot\,;\theta_*)H_Z^0(\cdot)$.

PROOF OF THEOREM 6.3: Under H_0, it follows from Lemma 6.2 that $f(\cdot\,|\,\cdot\,;\theta_*) = g(\cdot\,|\,\cdot\,;\gamma_*)$. Then, Part (i) follows from Theorem 3.3-(i), the continuity of $M_{p+q}(x;\lambda)$ 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.

PROOF OF LEMMA 7.1: 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 Assumptions A1-(b) and A8, $g(\cdot\,|\,\cdot\,;\tilde{\gamma}) = f(\cdot\,|\,\cdot\,;\theta_*)$ for H_X^0-almost all (y, z), which implies $E^0[\log g(Y_t|Z_t;\tilde{\gamma})] \geqslant E^0[\log f(Y_t|Z_t;\theta)]$ for any θ in Θ and, in particular for any $\theta = \phi(\gamma)$ for $\gamma \in \Gamma$. Using again Assumption A8, we have $E^0[\log g(Y_t|Z_t;\tilde{\gamma})] \geqslant E^0[\log g(Y_t|Z_t;\gamma)]$ 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 Assumptions A1-(b) and A8.

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

(iii) $\Rightarrow$ (ii): Suppose that $\theta_* \notin \phi(\Gamma)$; then $\theta_* \neq \tilde{\theta} \equiv \phi(\gamma_*)$. From (iii), Assumptions A1-(b) and 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_*.

LEMMA B: Given Assumptions A1-(b), A2–A5 and A8, we have under H_0^θ:

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

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

$$\text{(iii)} \qquad q \leqslant p, \quad \operatorname{rank} \frac{\partial\phi'(\gamma_*)}{\partial\gamma} = q.$$

PROOF OF LEMMA B: Under Assumptions A1-(b) and A8, $\partial \log g(\cdot \mid \cdot\,; \gamma)/\partial\gamma = \partial\phi'/\partial\gamma \cdot \partial \log f(\cdot \mid \cdot\,; \phi(\gamma))/\partial\theta H_X^0$-almost surely. But under H_0^θ, we have $\theta_* = \phi(\gamma_*)$ (Lemma 7.1), which establishes (ii) and the first equality of (i) using the definitions of B_g, B_f, and B_{gf}. In addition:

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

H_X^0-almost surely, where we have omitted the arguments of the functions, and where ϕ_k is the kth component of ϕ. Since $E^0[\partial \log f(Y_t \mid Z_t; \theta_*)/\partial\theta] = 0$ and since $\theta_* = \phi(\gamma_*)$, the second equality of (i) follows. Finally, (iii) follows from this equality and the fact that $A_f(\theta_*)$ and $A_g(\gamma_*)$ are nonsingular matrices (see, White (1982a), Theorem 3.1)).

PROOF OF THEOREM 7.2: Since under H_0^θ, we have $f(\cdot \mid \cdot\,; \theta_*) = g(\cdot \mid \cdot\,; \gamma_*)$ (Lemma 7.1), then (i) follows from (7.3) and Theorem 3.3-(i) if we show that the nonzero eigenvalues λ_* of W are the nonzero eigenvalues of $\underline{W}$. But, using Lemma B, the eigenvalues of W solve:

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

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

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

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 $\partial\phi'/\partial\gamma$ (Lemma B-(iii)), and where the third equation follows from the second equation by adding to the first-column matrices the second-column matrices postmultiplied by $-\partial\phi'/\partial\gamma$. Hence, the eigenvalues of W solve:

$$(A.23) \qquad 0 = \lambda^q \det\left\{ -B_f A_f^{-1} + B_f \frac{\partial\phi}{\partial\gamma'} A_g^{-1} \frac{\partial\phi'}{\partial\gamma} - \lambda I_p \right\},$$

which establishes that the nonzero eigenvalues of W are the nonzero eigenvalues of $\underline{W}$ as defined by (7.4). Equation (A.23) also shows that the eigenvalues of $\underline{W}$ are all real and nonnegative since $A_f^{-1} - [\partial\phi/\partial\gamma']A_g^{-1}[\partial\phi'/\partial\gamma] = A_f^{-1} - [\partial\phi/\partial\gamma']([\partial\phi'/\partial\gamma]A_f[\partial\phi/\partial\gamma'])^{-1}[\partial\phi'/\partial\gamma]$ which is n.s.d. Part (ii) follows from Lemma 3.1 and $H_A^\theta = H_f$.

PROOF OF COROLLARY 7.3: If $A_f + B_f = 0$, then it follows from Lemma B-(i) that under H_0^θ, $A_g + B_g = 0$. Part (i) follows from Corollary 3.4. Part (ii) is identical to Theorem 7.2-(ii).

PROOF OF THEOREM 7.4: Since $H_0^\theta = H_0^\omega$, (i) follows from Theorem 4.3 since the nonzero eigenvalues of W are the eigenvalues of $\underline{W}$. Part (ii) follows from Lemma 4.2 since H_A^θ is equivalent to H_A^ω. Part (iii) is proved similarly.

PROOF OF COROLLARY 7.5: As noticed in the proof of Corollary 7.3, both information matrix equivalences hold under H_0^θ. Then (i) follows from Corollary 4.4-(iv) since $B_{gf} B_f^{-1} B_{fg} B_g^{-1}$ is equal to I_q (using Lemma B). Parts (ii) and (iii) are identical to Theorem 7.4-(ii) and (iii).

REFERENCES

AGUIRRE-TORRES, V., AND A. R. GALLANT (1983): "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, 5–33.

AKAIKE, H. (1973): "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, 257–281.

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

AMEMIYA, T. (1980): "Selection of Regressors," *International Economic Review*, 21, 331–354.

ATKINSON, A. C. (1969): "A Test for Discriminating between Models," *Biometrika*, 56, 337–347.

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

BAUER, H. (1972): *Probability Theory and Elements of Measure Theory*. New York: Holt, Rinehart, and Winston.

BURGUETTE, J., A. R. GALLANT, AND G. SOUZA (1982): "On the Unification of the Asymptotic Theory of Nonlinear Econometric Models," *Econometric Reviews*, 1, 151–190.

CHOW, G. (1980): "The Selection of Variables for Use in Prediction: A Generalization of Hotelling's Solution," *Quantitative Econometrics and Development*, ed. by L. N. Klein, M. Nerlove, and S. C. Tiang. New York: Academic Press.

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

——— (1983): *Econometrics*. New York: McGraw-Hill, 1983.

COX, D. R. (1961): "Tests of Separate Families of Hypotheses," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 105–123.

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

DAVIDSON, R., AND J. G. MACKINNON (1981): "Several Tests for Model Specification in the Presence of Alternative Hypotheses," *Econometrica*, 49, 781–793.

DUBIN, J., AND D. RIVERS (1986): *Statistical Software Tools*. Pasadena: California Institute of Technology.

FOUTZ, R. V., AND R. C. SRIVASTAVA (1977): "The Performance of the Likelihood Ratio Test When the Model is Incorrect," *Annals of Statistics*, 5, 1183–1194.

GOURIEROUX, C., A. MONFORT, AND A. TROGNON (1983): "Testing Nested or Non-Nested Hypotheses," *Journal of Econometrics*, 21, 83–115.

——— (1984): "Pseudo Maximum Likelihood Method: Theory," *Econometrica*, 52, 681–700.

HENDRY, D. F. (1983): "Comment," *Econometric Reviews*, 2, 111–114.

HORN, R. G., AND C. A. JOHNSON (1985): *Matrix Analysis*. Cambridge: Cambridge University Press.

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

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

JOHNSON, N. L., AND S. KOTZ (1969): "Tables of Distributions of Positive Definite Quadratic Forms in Central Normal Variables," *Sankhya*, Series B, 303–314.

——— (1970): *Continuous Univariate Distributions-2*. New York: John Wiley and Sons.

JUDGE, G. G., W. E. GRIFFITHS, R. C. HILL, H. LUTKEPOHL, AND T. C. LEE (1985): *The Theory and Practice of Econometrics*. New York: John Wiley and Sons, Second edition.

KENT, J. T. (1982): "Robust Properties of Likelihood Ratio Tests," *Biometrika*, 69, 19–27.

KULLBACK, S., AND R. A. LEIBLER (1951): "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22, 79–86.

LIEN, D., AND Q. H. VUONG (1987): "Selecting the Best Linear Regression Model: A Classical Approach," *Journal of Econometrics*, *Annals*, 35, 3–23.

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

MCALEER, M., AND A. BERA (1983): "Comment," *Econometric Reviews*, 2, 121–130.

MIZON, G., AND J. F. RICHARD (1986): "The Encompassing Principle and its Application to Non-Nested Hypotheses," *Econometrica*, 54, 657–678.

MONFORT, A. (1980): *Cours de Probabilites*. Paris: Economica.

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

NEYMAN, J., AND E. S. PEARSON (1928): "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," *Biometrika*, 20A, 175–240.

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

——— (1987): "Global and Partial Non-Nested Hypotheses and Asymptotic Local Power," *Econometric Theory*, 3, 69–97.

PESARAN, M. H., AND A. S. DEATON (1978): "Testing Non-Nested Nonlinear Regression Models," *Econometrica*, 46, 677–694.

RAO, C. R. (1973): *Linear Statistical Inference and its Applications*. New York: John Wiley and Sons.

RAO, C. R., AND S. K. MITRA (1971): *Generalized Inverse of Matrices and its Applications*. New York: John Wiley and Sons.

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

SCHWARZ, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

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

——— (1986): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," Social Science Working Paper, No. 605. Pasadena: California Institute of Technology.

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

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

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

——— (1983): "Editor's Introduction," *Journal of Econometrics*, 21, 1–3.

WHITE, H., AND I. DOMOWITZ (1984): "Non-linear Regression with Dependent Observations," *Econometrica*, 52, 143–161.

WHITE, H., AND L. OLSON (1979): "Determinants of Wage Change on the Job: A Symmetric Test of Non-Nested Hypotheses," mimeo, University of Rochester.

WILKS, S. S. (1938): "The Large Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses," *Annals of Mathematical Statistics*, 9, 60–62.