DISCREPANCY RISK MODEL SELECTION TEST THEORY FOR COMPARING POSSIBLY MISSPECIFIED OR NONNESTED MODELS

R.M. GOLDEN

APPLIED COGNITION AND NEUROSCIENCE PROGRAM
UNIVERSITY OF TEXAS AT DALLAS

A new model selection statistical test is proposed for testing the null hypothesis that two probability models equally effectively fit the underlying data generating process (DGP). The new model selection test, called the Discrepancy Risk Model Selection Test (DRMST), extends previous work (see Vuong, 1989) on this problem in four distinct ways. First, generalized goodness-of-fit measures (which include log-likelihood functions) can be used. Second, unlike the classical likelihood ratio test, the models are not required to be fully nested where the nesting concept is defined for generalized goodness-of-fit measures. The DRMST also differs from the likelihood ratio test by not requiring that either competing model provides a completely accurate representation of the DGP. And, fourth, the DRMST may be used to compare competing time-series models using correlated observations as well as data consisting of independent and identically distributed observations.

Key words: asymptotic statistical theory, model selection, hypothesis-testing, model misspecification, time-series, m-estimation

Considerable research in the field of model selection has focussed upon the Model Selection Criterion (MSC) problem where multiple competing descriptions of some Data Generating Process (DGP) are compared (e.g., Akaike, 1973; Balasubramanian, 1997; Bozdogan, 1987; Clarke & Barron, 1990; Djuric, 1998; Kass & Wasserman, 1995; Linhart & Zuchhini, 1986; Myung, Forster, & Browne, 2000; Qian & Kunsch, 1998; Rissanen, 1996; Schwarz, 1978). In this paradigm, some goodness-of-fit measure is used to estimate the (true) expected goodness-of-fit of each *model* (i.e., a family of probability distributions) to the underlying DGP. The model (or models) which has (have) the smallest *estimated* goodness-of-fit is (are) then selected.

This paper focuses attention upon a closely related and equally important problem called the Model Selection Test (MST) problem which has received relatively less attention in the literature. In this paradigm, the standard error of the difference in the relative fits of the two models is estimated and used to test the null hypothesis that both models provide equally effective fits to the underlying DGP at a chosen significance level.

An important early approach to the MST problem is the Generalized Likelihood Ratio Test (GLRT) which is now widely used (Wilks, 1938). The essential strategy for applying the GLRT is to compute maximum likelihood estimates of the model's parameters using the observed data. The goodness-of-fit of the model is then assessed by using the computed maximum likelihood estimates to calculate the observed likelihood of the data given the "full" model. It is also assumed that the DGP is a probability distribution contained in the full model. Some subset of the model's parameters are then usually set equal to a constant (such as zero) and another maximum likelihood estimation procedure is then used to estimate the free parameters of the "restricted"

model. Thus, it is assumed in the GLRT that the two competing models satisfy a "nesting" relationship where the restricted model is nested within the full model. The GLRT is then used to test the null hypothesis that the observed difference in the log-likelihood goodness-of-fit for the two models is due to chance.

In this paper, a new MST will be proposed which relaxes several key assumptions of the GLRT. First, generalized goodness-of-fit measures (which include log-likelihood goodness-of-fit measures as special cases) will be considered. Second, the two competing models are not required to satisfy a nesting relationship. And third, it is not required that either model be correctly specified (i.e., both models may be inadequate for representing the DGP; see Foutz & Srivastava, 1977; Golden, 1995, 1996, 2000; Vuong, 1989; White 1982, 1984 for further discussion of this issue). Moreover, the new MST is applicable to time-series data as well as data which satisfies an independent and identically distributed (i.i.d.) assumption.

Historically, both Linhart (1988; also see Efron, 1984) and Vuong (1989) showed how the null hypothesis that two *strictly nonnested models* (i.e., two models which have no probability distribution in common) provide equally effective descriptions of the DGP could be tested. Problems in applying this methodology arose however when the models were not strictly nonnested (see Shimodaira, 1997). Vuong addressed this problem using a two stage MST applicable to situations where the two competing models could be both misspecified and satisfy any type of nesting relationship (e.g., strictly nonnested, fully nested, partially nested, etc.) for the case where the goodness-of-fit measure was a log-likelihood function and the observations were i.i.d. (i.e., independent and identically distributed). Vuong and Wang (1993) considered a MST closely related to the original Vuong (1989) methodology which was applicable to the i.i.d. observation case for a least-squares type goodness-of-fit function. More recently, Golden (2000) proposed a generalized version of Vuong's (1989) methodology for the i.i.d. case. From an empirical perspective, Vuong's (1989) model selection theory has been successfully applied in a variety of task domains (for some recent applications see Barth, Beaver, & Landsman, 1998; Chen & Plott, 1998; Gertham, 1997; Golden, 1995, 1998; Grootendorst, 1995; Wang, Halbrendt, & Johnson, 1996; Vincent, 1999).

Using the methods of Vuong (1989) and White (1994), this paper generalizes the development of DRMST theory for the i.i.d. case described by Golden (2000) to a fairly general time-series analysis case. Although one of the main results of this paper, Theorem 4 (Appendix), is in fact a special case of an important analysis[1] by Rivers and Vuong (2002), the statement and proof of Theorem 4 is still distinguished by its underlying modeling assumptions. These modeling assumptions are substantially easier to understand and verify relative to the Rivers and Vuong approach.

This paper is organized in the following manner. First, the DRMST will be introduced in conjunction with some essential notation. Second, the formal assumptions of the DRMST theory are presented and discussed. Key theorems and proofs are presented in the Appendix. These theorems are required to establish asymptotic upper bounds on the Type 1 error probability and conditions for the Type 2 error probabilities to converge to zero. The theorems are also used to establish the consistency of the various estimators used in this paper.

1. The Discrepancy Risk Model Selection Test

It will be assumed that the observed data is a particular realization of a strictly stationary Data Generating Process denoted by the notation DGP. The assumption that the DGP is strictly stationary implies that all observations are identically distributed yet they may be highly correlated in time. Thus, the DRMST is applicable to a large class of time-series data analysis problems. The DGP is a probability distribution which is denoted by a dot in Figure 1. A proba-
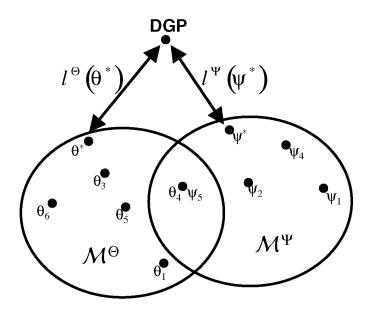
---

[1]Theorem 4 in the Appendix was developed independently of knowledge of the recently published paper by Rivers and Vuong (2002).

FIGURE 1.

Discrepancy Risk Model Selection Test Null Hypothesis. Probability distributions in this figure are depicted as black dots while the two probability models in this figure, $\mathcal{M}^{\Theta}$ and $\mathcal{M}^{\Psi}$, specify specific sets of probability distributions. The minimum expected discrepancy loss (where the expectation is taken with respect to the distinguished DGP probability distribution) for probability model $\mathcal{M}^{\Theta}$ is denoted by $l^{\Theta}(\theta^*)$. Similarly, the minimum expected discrepancy loss for $\mathcal{M}^{\Psi}$ is denoted by $l^{\Psi}(\psi^*)$ where the expectation is also taken with respect to the DGP. The DRMST null hypothesis is $H_0 : l^{\Theta}(\theta^*) = l^{\Psi}(\psi^*)$.

bility model is a set of probability distributions whose elements may be indexed by a "parameter vector." Figure 1 depicts two probability models

$$\mathcal{M}^{\Theta} = \{p_{\theta_1}, p_{\theta^*}, p_{\theta_3}, p_{\theta_4}, p_{\theta_5}, p_{\theta_6}\} \quad \text{and} \quad \mathcal{M}^{\Psi} = \{p_{\psi_1}, p_{\psi_2}, p_{\psi^*}, p_{\psi_4}, p_{\psi_5}\}$$

which consist of six and five probability distributions respectively. The "parameter space" Θ for $\mathcal{M}^{\Theta}$ is a subset of a p-dimensional real vector space defined such that $\theta_j \in \Theta$. Similarly, the parameter space Ψ for $\mathcal{M}^{\Psi}$ is a subset of a q-dimensional real vector space defined such that $\psi_j \in \Psi$. Figure 1 also depicts the *discrepancy risk functions* l^{Θ} and l^{Ψ} evaluated at $\theta^* = $ argmin l^{Θ} and $\psi^* = $ argmin l^{Ψ} respectively. In particular, $l^{\Theta}(\theta^*)$ is the expected loss associated with selecting the best-fitting probability distribution, p_{θ^*}, from $\mathcal{M}^{\Theta}$. The goal of this article is to develop a model selection test (i.e., the DRMST) in order to test the null hypothesis $H_0 :$ $l^{\Theta}(\theta^*) = l^{\Psi}(\psi^*)$ that is, models $\mathcal{M}^{\Theta}$ and $\mathcal{M}^{\Psi}$ provide equally effective fits to the DGP.

Referring to Figure 1, note that $\mathcal{M}^{\Theta} \cap \mathcal{M}^{\Psi} = \{p_{\theta_4}\} = \{p_{\psi_5}\}$. That is, the two probability models are "overlapping" or "nonnested" since neither probability model is a subset of the other. Also note that the DGP is not contained in either probability model indicating that both probability models are *misspecified* with respect to the DGP. In the classic Generalized Likelihood Ratio Test it is assumed that one probability model (known as the "reduced model") is entirely contained within the other probability model (known as the "full model") and that the DGP is an element of the full model. As noted in the previous section, the DRMST relaxes the two assumptions of fully nested and correctly specified models considerably. And thus, provides a mechanism for the analysis of nested, nonnested, and misspecified models of stationary time-series data within a unified framework.

A two-stage MST is used to test the DRMST null hypothesis as illustrated in Figure 2. In the first stage, a special statistical test called the Variance MST is done. If the null hypothesis for the Variance MST is accepted, then the DRMST null hypothesis is accepted. If the null hypothesis for
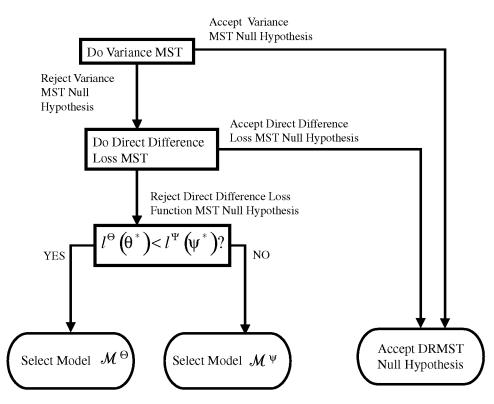
FIGURE 2.

The Discrepancy Risk Model Selection Test. The null hypothesis for the DRMST is that the two competing models have exactly the same expected discrepancy loss. First, the Variance Model Selection Test is done. If the null hypothesis for the Variance MST is accepted, the DRMST null hypothesis is accepted. If the null hypothesis for the Variance MST is rejected, then apply the Direct Difference Loss MST. If the null hypothesis for the Direct Difference Loss MST is accepted, accept the DRMST null hypothesis. If the null hypothesis for the Direct Difference Loss MST is rejected, select the model whose estimated expected discrepancy loss is smallest.

the Variance MST is rejected, then the second stage MST called the Direct Difference Loss MST is done. If the null hypothesis for the Direct Difference Loss MST is accepted, then the DRMST null hypothesis is accepted. If the null hypothesis for the Direct Difference Loss MST is rejected, then the DRMST null hypothesis is rejected. Given the DRMST null hypothesis is rejected, then the model with the smallest estimated expected loss is selected. Theorem 5 of the Appendix summarizes the results of the theoretical analyses and establishes minimal necessary conditions for the DRMST to be a legitimate statistical test. Specifically, the Type 2 error probability approaches zero as the sample size increases and the Type 1 error probability is asymptotically bounded from above by the DRMST significance level.

1.1. Parameter Estimation

The parameter estimates for both competing models are obtained by minimizing appropriate estimators of each model's discrepancy risk function. The first parameter estimate (denoted by $\theta_n^{\&}$) is defined as the strict global minimum of the *sample discrepancy risk function*, $l_n^{\Theta} : \Theta \times \Gamma^n \to \mathcal{R}$,

$$l_n^{\Theta}(\theta, \mathbf{X}_n) = \left(\frac{1}{n}\right) \sum_{i=1}^{n} c^{\Theta}(\theta, \mathbf{x}_i)$$

where the notation $\mathbf{X}_n = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \Gamma^n \subseteq \mathcal{R}^{dn}$ denotes the observed data. The function c^Θ is referred to as the *discrepancy loss function*.

The second type of parameter estimate $\theta_n^\#$ is defined as the strict global minimum of the *model selection criterion* function, $\mathcal{C}_n^\Theta = l_n^\Theta + k_n^\Theta$ where $k_n^\Theta : \Theta \times \Gamma^n \to \mathcal{R}$ is called the *penalty term*. It is always assumed that the penalty term k_n^Θ has the property that: $\sqrt{n}k_n^\Theta \to 0$ in probability as $n \to \infty$.

These two slightly different parameter estimation goals are required in order to accommodate parameter estimation methods which compute $\theta_n^\#$ (such as Bayesian Ridge Regression as described by Draper & Smith, 1981, p. 319) as well as methods which compute $\theta_n^\&$ (such as the Asymptotic MAP Criteria described by Djuric, 1998). In the following discussion, and throughout the remainder of this paper, the convention will be used that all results which hold for both $\theta_n^\#$ or $\theta_n^\&$ will be stated in terms of the quantity θ_n^*. Moreover, the notation θ_n^* and ψ_n^* will be used to refer to the parameter estimates associated with the models $\mathcal{M}^\Theta$ and $\mathcal{M}^\Psi$ respectively.

1.1.1. Maximum Likelihood Estimation

In the case of i.i.d. observations, an important choice of the functional form of c^Θ (considered in the analysis by Vuong, 1989) is to choose c^Θ to be a *log-likelihood loss function* so that $c^\Theta(\theta, \mathbf{x}_i) = -\log(p(\mathbf{x}_i \mid \theta))$ where $p(\mathbf{x}_i \mid \theta)$ is the likelihood of observation $\mathbf{x}_i$ given parameter vector θ. Now consider the time-series case involving correlated and identically distributed observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots)$ which are presumed to be generated by a k-th order Markov probability law of the form $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-k}; \theta)$. The log-likelihood loss function for this stochastic process is obtained by defining c^Θ such that

$$c^\Theta(\theta, \mathbf{y}_t) = -\log\left[p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-k}; \theta)\right]$$

with $\mathbf{y}_t = [\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-k}]$.

1.1.2. Defining the Parameter Space

The DRMST theory assumes that the model parameter space, Θ, is a convex, closed, and bounded subset of $\mathcal{R}^p$. It is also assumed that the objective function l_n^Θ has a unique minimum (i.e., a strict global minimum) on the model parameter space. If the log-likelihood discrepancy risk function is convex over a large region of the parameter space (as is the case for linear exponential models), then it is reasonable to define Θ to be very large.

On the other hand, suppose the discrepancy risk function has multiple strict local minima and/or strict global minima. This latter situation frequently arises in nonlinear regression modeling and artificial neural network modeling (e.g., Golden, 1995; White, 1989). In this case, simply choose Θ to be a sufficiently small closed, bounded convex region which contains exactly one strict local minimum in $\mathcal{R}^p$ which happens to be the unique global minimum on Θ. Such a formulation of the problem also provides a mechanism for using the DRMST to test the null hypothesis that two given strict local minima are "equally deep" (see Gan & Jian, 1999, for a related analysis and discussion).

1.1.3. Checking for Convergence of Parameter Estimation Algorithm

The parameter estimates are defined as the solution to an equation of the form: $\nabla \mathcal{C}_n^\Theta$ or ∇l_n^Θ depending upon the particular application. Accordingly, an appropriate norm of the gradient vector (e.g., the element of the gradient vector with the largest absolute value)

$$\nabla l_n^\Theta(\theta_n^*, \mathbf{X}_n) = \left(\frac{1}{n}\right) \sum_{i=1}^{n} \nabla c^\Theta(\theta_n^*, \mathbf{x}_i) \tag{1}$$

should be computed and checked to confirm the gradient norm is sufficiently close to zero. Verification of this condition indicates that the parameter estimation algorithm has properly terminated upon a critical point of the objective function l_n^{Θ}.

1.1.4. Checking Uniqueness of the Parameter Estimates

Insights into whether or not a random variable which takes on the value of the critical point, θ_n^*, of l_n^{Θ} is converging to a locally unique solution, θ^*, as the sample size n increases may be obtained by inspecting the statistic $\mathbf{A}_n^{\Theta} = \nabla^2 l_n^{\Theta}$ evaluated at θ_n^*.

The matrix function $\mathbf{A}_n$ is defined by the formula

$$\mathbf{A}_n = \begin{bmatrix} \mathbf{A}_n^{\Theta} & \mathbf{0}_{p,q} \\ \mathbf{0}_{q,p} & \mathbf{A}_n^{\Psi} \end{bmatrix}$$

where $\mathbf{0}_{p,q}$ is a $p \times q$ dimensional matrix of zeros, $\mathbf{A}_n^{\Theta} = \nabla^2 l_n^{\Theta}$, and $\mathbf{A}_n^{\Psi} = \nabla^2 l_n^{\Psi}$. The matrix $\mathbf{A}_n^*$ is obtained by evaluating $\mathbf{A}_n$ at $\omega_n^* = (\theta_n^*, \psi_n^*)$.

If all eigenvalues of $\mathbf{A}_n^*$ are strictly positive, then this indicates the critical point θ_n^* is a strict local minimum of l_n^{Θ} and ψ_n^* is a strict local minimum of l_n^{Ψ}. The condition number of $\mathbf{A}_n^*$ (defined by the ratio of the largest to smallest eigenvalue of $\mathbf{A}_n^*$) is then computed. A large condition number provides an indication that as the sample size becomes large the critical point may not possibly correspond to a *strict* local minimum of the objective function (i.e., be"locally unique"). Computation of $\mathbf{A}_n^*$ for several choices of n (i.e., several subsets of the data sample) may be helpful in obtaining insight into the issue of whether or not $\mathbf{A}_n^*$ is converging in fact to a positive definite matrix $\mathbf{A}^*$ (as opposed to a singular matrix). Note that this analysis is analogous to checking for the presence of multicollinearity in the special case of linear regression.

1.2. Checking a Necessary Condition of the Asymptotic Approximations

An important analysis which is necessary but not sufficient to establish that the asymptotic approximations provided in this paper hold involves examining another matrix called $\mathbf{B}_n^*$ in a manner analogous to the way $\mathbf{A}_n^*$ was examined. The construction of the relevant statistics requires the use of some broad assumptions regarding the DGP. Although these assumptions will be discussed in greater detail later, for now it is enough to note that one must make an assumption regarding the statistical independence of the members of the stochastic process generated by the DGP.

Specifically, if all elements of the DGP are statistically independent we say the DGP is a τ-dependent process with $\tau = 0$. More generally, denoting the DGP as the stochastic sequence of observations $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots$, then we say the DGP is τ-dependent if there exists a finite nonnegative integer constant τ such that $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$ are statistically independent for all $|i - j| > \tau$. Thus, a DGP with $\tau = 3$ can also be described as a DGP with $\tau = 4$ or even a DGP with $\tau = 10000$. Thus, in situations where the exact value of τ is unknown, a conservative strategy might involve choosing larger as opposed to smaller values of τ.

Define $J_{i,\tau} = \{m \in \{1, 2, 3, \ldots, n\} : |m - i| \leq \tau\}$. Let c_i^{Θ} denote the discrepancy loss function c^{Θ} for model $\mathcal{M}^{\Theta}$ evaluated at observation $\mathbf{x}_i$. That is, $c_i^{\Theta} = c^{\Theta}(\boldsymbol{\theta}, \mathbf{x}_i)$. The random variable $\tilde{c}_i^{\Theta}$ is similarly defined by the formula $\tilde{c}_i^{\Theta} = c^{\Theta}(\boldsymbol{\theta}, \tilde{\mathbf{x}}_i)$. The quantities c_i^{Ψ} and $\tilde{c}_i^{\Psi}$ are defined in a manner analogous to the definitions of c_i^{Θ} and $\tilde{c}_i^{\Theta}$ but with respect to the competing model $\mathcal{M}^{\Psi}$.

Now define the matrix function

$$\mathbf{B}_n^{\Theta} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} \sum_{j \in J_{i,\tau}} [\nabla c_i^{\Theta}][\nabla c_j^{\Theta}]^T,$$

define the matrix function

$$\mathbf{B}_n^\Psi = \left(\frac{1}{n}\right) \sum_{i=1}^{n} \sum_{j \in J_{i,\tau}} [\nabla c_i^\Psi][\nabla c_j^\Psi]^T,$$

and define the matrix function

$$\mathbf{B}_n^{\Theta,\Psi} = [\mathbf{B}_n^{\Psi,\Theta}]^T = \left(\frac{1}{n}\right) \sum_{i=1}^{n} \sum_{j \in J_{i,\tau}} \nabla c_i^\Theta [\nabla c_j^\Psi]^T.$$

The matrix function $\mathbf{B}_n$ is then defined by the formula

$$\mathbf{B}_n = \begin{bmatrix} \mathbf{B}_n^\Theta & \mathbf{B}_n^{\Theta,\Psi} \\ \mathbf{B}_n^{\Psi,\Theta} & \mathbf{B}_n^\Psi \end{bmatrix}.$$

Evaluating the matrix function $\mathbf{B}_n$ at the point $\omega_n^* = (\theta_n^*, \psi_n^*)$ yields the random matrix $\mathbf{B}_n^*$. Theorem 1 of the Appendix shows that $\mathbf{B}_n^*$ converges with probability one to a constant matrix $\mathbf{B}^*$. This observation suggests a practical mechanism for checking a key assumption of the DRMST Theory (see Assumption A11 in sec. 2) which is that $\mathbf{B}^*$ must be positive definite. To check this assumption, the condition number of $\mathbf{B}_n^*$ may be checked in a manner similar to the analysis of $\mathbf{A}_n^*$. If the condition number of $\mathbf{B}_n^*$ does not appear to be the realization of an observation in a stochastic sequence converging to a finite positive number, then this indicates Assumption A11 may not be true. If Assumption A11 is false, then the DRMST theory results may not be valid.

1.3. Checking the Discrepancy Autocorrelation Coefficient

If it is assumed that the observations are not independent, then it is helpful to examine a specially-defined correlation coefficient which is the estimated expected average correlation between $\tilde{c}_t^\Theta - \tilde{c}_t^\Psi$ and $\tilde{c}_{t+k}^\Theta - \tilde{c}_{t+k}^\Psi$ (for $k = 1, 2, \ldots$). This special correlation coefficient is called the *estimated discrepancy autocorrelation coefficient* and is defined by the formula $r_n = \bar{r}(\omega_n^*, \mathbf{X}_n)$ where

$$\bar{r} = \frac{\left[\sum_{i=1}^{n} \sum_{j \in J_{i,\tau}, j \neq i} (c_i^\Theta - c_i^\Psi)(c_j^\Theta - c_j^\Psi) \right]}{2\tau \sum_{i=1}^{n} (c_i^\Theta - c_i^\Psi)^2}.$$

Define θ^* and ψ^* as the respective unique global minima of the discrepancy risk functions, l^Θ, and l^Ψ. Define $\omega^* = [\theta^*, \psi^*]$. The *true discrepancy autocorrelation coefficient* is defined by the formula $r^* = \bar{r}^*(\omega^*) = \bar{r}^*([\theta^*, \psi^*])$ where

$$\bar{r}^* = \frac{\sum_{j=1}^{\tau} E\{(\tilde{c}_i^\Theta - \tilde{c}_i^\Psi)(\tilde{c}_{i+j}^\Theta - \tilde{c}_{i+j}^\Psi)\}}{\tau E\{(\tilde{c}_i^\Theta - \tilde{c}_i^\Psi)^2\}} \tag{2}$$

where all expectations are taken with respect to $\tilde{\mathbf{X}}_n$.

In situations where the DGP does not consist of independent observations, the DRMST theory developed in this article assumes that $r^* \neq -1/(2\tau)$. Specifically, the asymptotic distribution of the DRMST statistic is derived under the assumption that $r^* \neq -1/(2\tau)$. In order to estimate r^*, the number r_n may be computed from the sample since r_n is an asymptotically unbiased estimate of r^* as shown in Theorem 1 of the Appendix. One practical approach to checking the condition that $r^* \neq -1/(2\tau)$ would be to compute r_n for different subsamples of different sample sizes from the data in order to determine if r_n appears to be converging to the number $-1/(2\tau)$.

The meaning of the condition $r^* \neq -1/(2\tau)$ can be clarified by defining the function $\sigma_Z^2 : (\Theta \times \Psi) \to [0, \infty)$ such that

$$\sigma_Z^2 = E\{(\tilde{c}_i^{\Theta} - \tilde{c}_i^{\Psi})^2\} + 2 \sum_{j=1}^{\tau} E\{(\tilde{c}_i^{\Theta} - \tilde{c}_i^{\Psi})(\tilde{c}_{i+j}^{\Theta} - \tilde{c}_{i+j}^{\Psi})\} \tag{3}$$

where all expectations are taken with respect to $\tilde{\mathbf{X}}_n$. In the special case where $\tau = 0$ (i.e., the observations are independent), then the second term in (3) vanishes, and it follows that the null hypothesis

$$H_0 : \sigma_{Z^*}^2 = \sigma_Z^2(\theta^*, \psi^*) = 0$$

and the null hypothesis

$$H_0^{\sigma^2}(\tau) : c^{\Theta}(\theta^*, \tilde{\mathbf{x}}_t) = c^{\Psi}(\psi^*, \tilde{\mathbf{x}}_t) \quad \text{w.p.1}$$

are equivalent.

The case where τ is a positive integer is now considered. Combining (3) and (2) it follows that

$$\sigma_{Z^*}^2 = (1 + 2\tau r^*) E\{(\tilde{c}_i^{\Theta} - \tilde{c}_i^{\Psi})^2\}. \tag{4}$$

Inspection of (4) shows that by assuming $r^* \neq -1/(2\tau)$ (i.e., $(1 + 2\tau r^*) \neq 0$) it follows that the null hypothesis $H_0 : \sigma_{Z^*}^2 = 0$ and the null hypothesis $H_0^{\sigma^2}(\tau)$ are equivalent. Thus, a statistical test designed to test the null hypothesis that $\sigma_{Z^*}^2 = 0$ may be used to determine if one model fits the observed data significantly better than another model. This new statistical test will be referred to as the "Variance MST" and is discussed in the next section.

1.4. The Variance Model Selection Test

As noted at the end of the previous section, the Variance MST is designed to test the null hypothesis $H_0^{\sigma^2}(\tau) : c^{\Theta}(\theta^*, \tilde{\mathbf{x}}_t) = c^{\Psi}(\psi^*, \tilde{\mathbf{x}}_t)$ w.p.1. Semantically, the null hypothesis $H_0^{\sigma^2}(\tau)$ specifies a state of the world where the null hypothesis for the DRMST is accepted as well. The Variance MST is generally quite computationally intensive and can be avoided if one can demonstrate analytically that $H_0^{\sigma^2}(\tau)$ is false. Linhart (1988), for example, simply used the Direct Difference Loss MST without using the Variance MST since he considered two log-likelihood loss functions where the two competing probability models were a lognormal model and a gamma model.

In many commonly arising situations, however, the Variance MST can not be avoided. For example, consider the situation where the loss functions have the form of negative log-likelihood loss functions. Also assume that the Variance MST is being used to compare two probability models where one probability model is fully nested within the other. If the reduced model contains the optimal distribution corresponding to the DGP, then that same optimal distribution will be obtained from parameter estimation on the full model. Thus, the general case where the models are fully nested and the null hypothesis is true corresponds to a situation where $c^{\Theta}(\theta^*, \tilde{\mathbf{x}}_t) = c^{\Psi}(\psi^*, \tilde{\mathbf{x}}_t)$ w.p.1. The Variance MST can not generally be avoided in such situations.

The calculations required for the Variance MST are now provided. First, compute the *discrepancy variance*, $(\sigma_{Z_n}^*)^2 = \sigma_{Z_n}^2(\omega_n^*)$, given by the formula

$$\sigma_{Z_n}^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in J_{i,\tau}} \left(c_i^{\Theta} - c_i^{\Psi} \right) \left(c_j^{\Theta} - c_j^{\Psi} \right)$$

evaluated at ω_n^*.

Second, compute the $p + q$-dimensional $\mathbf{R}_n$ matrix function defined by the formula

$$\mathbf{R}_n = \begin{bmatrix} -\mathbf{B}_n^{\Theta}[\mathbf{A}_n^{\Theta}]^{-1} & -\mathbf{B}_n^{\Theta,\Psi}[\mathbf{A}_n^{\Psi}]^{-1} \\ \mathbf{B}_n^{\Psi,\Theta}[\mathbf{A}_n^{\Theta}]^{-1} & \mathbf{B}_n^{\Psi}[\mathbf{A}_n^{\Psi}]^{-1} \end{bmatrix}$$

where $\mathbf{B}_n^{\Theta}$, $\mathbf{A}_n^{\Theta}$ have dimension p and $\mathbf{B}_n^{\Psi}$, $\mathbf{A}_n^{\Psi}$ have dimension q. Evaluate $\mathbf{R}_n$ at ω_n^* to obtain the matrix $\mathbf{R}_n^*$. Then, use $\mathbf{R}_n^*$ to compute a $p + q$-dimensional vector $\mathbf{w}_n$ whose i-th element is the square of the i-th eigenvalue of $\mathbf{R}_n^*$.

Third, the probability that a weighted chi-square random variable with weight vector equal to $\mathbf{w}_n$ exceeds $n(\sigma_{Z_n}^*)^2$, $p(\tilde{\chi}^2(\mathbf{w}_n) > n(\sigma_{Z_n}^*)^2)$, is then computed. A *weighted chi-square random variable*, $\tilde{\chi}^2(\mathbf{w})$ with *weight* parameter vector $\mathbf{w}$, may be expressed by the formula: $\tilde{\chi}^2(\mathbf{w}) = \sum_{i=1}^{d} w_i \tilde{z}_i^2$ where $\tilde{z}_i$ are i.i.d. normally distributed random variables and $\mathbf{w} = [w_1, \ldots, w_d]$ is called the *weight* parameter. Thus, the usual chi-square random variable with d degrees of freedom would be represented using this notation by the quantity $\tilde{\chi}^2(\mathbf{1}_d)$ where $\mathbf{1}_d$ denotes a d-dimensional vector of ones. Computer software for evaluating the weighted chi-square random variable cumulative distribution is not generally available unlike computer software for evaluating the chi-square random variable cumulative distribution. Accordingly, a special case of the method of Sheil and O'Muircheartaigh (1977; see Davies, 1980, for an alternative method) is provided for evaluating the cumulative distribution function for the weighted chi-square random variable in terms of the cumulative distribution functions of chi-square random variables.

Following the approach of Sheil and O'Muircheartaigh (1977), let ϵ be the maximum absolute error permitted in the approximation, $\hat{p}_\epsilon(\tilde{\chi}^2(\mathbf{w}) > n(\sigma_{Z_n}^*)^2)$, of the probability $p(\tilde{\chi}^2(\mathbf{w}) > n(\sigma_{Z_n}^*)^2)$. Let $\beta = 0.90625 w_{\min}$, where $w_{\min}$ is a smallest element of $\mathbf{w}$. Let w_i be the i-th element of $\mathbf{w}$.

$$\mu_0 = \prod_{i=1}^{p+q} (\beta/w_i)^{1/2}, \quad \text{and} \quad \mu_k = k^{-1} \sum_{j=0}^{k-1} g_{k-j} \mu_j, k = 1, 2, \ldots N_\epsilon,$$

where

$$g_{k-j} = \left(\frac{1}{2}\right) \sum_{s=1}^{p+q} \left(1 - \left[\frac{\beta}{w_s}\right]\right)^{k-j}.$$

Sheil and O'Muircheartaigh (1977) show there exists a positive integer N_ϵ such that for a given positive number ϵ:

$$\left(1 - \sum_{k=0}^{N_\epsilon} \mu_k\right) p\left(\tilde{\chi}^2(\mathbf{1}_{p+q+2N_\epsilon}) \leq \frac{n(\sigma_{Z_n}^*)^2}{\beta}\right) < \epsilon.$$

Then, with

$$\hat{p}_\epsilon(\tilde{\chi}^2(\mathbf{w}_n) > n(\sigma_{Z_n}^*)^2) = 1 - \sum_{k=0}^{N_\epsilon} \mu_k p\left(\tilde{\chi}^2(\mathbf{1}_{p+q+2k}) \leq \frac{n(\sigma_{Z_n}^*)^2}{\beta}\right),$$

Sheil and O'Muircheartaigh (1977) show that the absolute approximation error is uniformly bounded by ϵ. That is,

$$\left| \hat{p}_\epsilon(\tilde{\chi}^2(\mathbf{w}) > n(\sigma_{Z_n}^*)^2) - p(\tilde{\chi}^2(\mathbf{w}) > n(\sigma_{Z_n}^*)^2) \right| < \epsilon.$$

Given a method for computing $\hat{p}_\epsilon(\tilde{\chi}^2(\mathbf{w}_n) > n(\sigma_{Z_n}^*)^2) \geq \alpha$, the null hypothesis of the Variance MST can be tested by the following procedure where α is the significance level of the test. If $\hat{p}_\epsilon(\tilde{\chi}^2(\mathbf{w}_n) > n(\sigma_{Z_n}^*))^2 \geq \alpha$, accept the Variance MST null hypothesis; otherwise reject the Variance MST null hypothesis.

As shown by Theorem 3 in the Appendix, the Type 1 error probability of the Variance MST (i.e., accepting the null hypothesis when the null hypothesis is false) has an asymptotic upper bound of α as the sample size becomes large. In addition, Theorem 3 also shows that the Type 2 error probability of the Variance MST (i.e., rejecting the null hypothesis when the null hypothesis is true) converges to zero as the sample size becomes large.

1.5. The Direct Difference Loss Model Selection Test

The null hypothesis for the Direct Difference Loss MST is identical to the null hypothesis of the DRMST (i.e., $H_0 : l^{\Theta}(\theta^*) = l^{\Psi}(\psi^*)$). However, unlike the DRMST, the Direct Difference Loss MST makes the additional assumption that the null hypothesis of the Variance MST is false.

Let α be the chosen significance level for the DRMST (e.g., $\alpha = 0.05$ or $\alpha = 0.01$). Compute

$$Z_n = \frac{\sqrt{n}\left(\mathcal{C}_n^{\Theta}(\theta_n^*) - \mathcal{C}_n^{\Psi}(\psi_n^*)\right)}{\sigma_{Z_n}^*}$$

where $\sigma_{Z_n}^*$ is the square root of the discrepancy variance computed in the Variance MST. It is assumed that $\sigma_{Z_n}^*$ is strictly positive since the Direct Difference Loss MST is only invoked if either: (i) the Variance MST null hypothesis has been rejected, or (ii) direct mathematical analysis shows $\sigma_{Z_n}^*$ must converge almost surely to a strictly positive real number (see sec. 1.3; also see Linhart, 1988).

Referring to Figure 2 and using the notation of section 1.3, if $p(\tilde{\chi}^2(1) > Z_n^2) \geq \alpha$, then accept the DRMST null hypothesis $H_0 : l^{\Theta}(\theta^*) = l^{\Psi}(\psi^*)$. If $p(\tilde{\chi}^2(1) > Z_n^2) < \alpha$, then reject the DRMST null hypothesis. Assuming the DRMST null hypothesis is rejected, select model $\mathcal{M}^{\Theta}$ if $\mathcal{C}_n^{\Psi}(\psi_n^*) > \mathcal{C}_n^{\Theta}(\theta_n^*)$ and model $\mathcal{M}^{\Psi}$ if $\mathcal{C}_n^{\Psi}(\psi_n^*) < \mathcal{C}_n^{\Theta}(\theta_n^*)$.

As shown by Theorem 4 in the Appendix, the Type 1 error probability of the Direct Difference Loss MST has an asymptotic upper bound of α as the sample size becomes large.

In addition, Theorem 4 also shows that the Type 2 error probability of the Variance MST converges to zero as the sample size becomes large.

1.6. The Discrepancy Risk Model Selection Test

The DRMST (Discrepancy Risk Model Selection Test) is designed to test the null hypothesis that the two competing models have exactly the same expected discrepancy loss. The Discrepancy Risk Model Selection Test (DRMST) is described in Figure 2. First, the Variance MST is done. If the null hypothesis for the Variance MST is accepted, the DRMST null hypothesis is accepted. If the null hypothesis for the Variance MST is rejected, then apply the Direct Difference Loss MST. If the null hypothesis for the Direct Difference Loss MST is accepted, accept the DRMST null hypothesis. If the null hypothesis for the Direct Difference Loss MST is rejected, select the model whose estimated expected discrepancy loss is smallest.

Although Figure 2 indicates that the DRMST null hypothesis is accepted if either the Variance MST or the Direct Difference Loss MST null hypotheses are accepted, the Variance MST and Direct Difference Loss MST null hypotheses have distinct interpretations. Accepting the null hypothesis for the Variance MST is equivalent to concluding $c^{\Theta}(\theta^*, \tilde{\mathbf{x}}_t) = c^{\Psi}(\psi^*, \tilde{\mathbf{x}}_t)$ w.p.1 while accepting the null hypothesis for the Direct Difference Loss MST is equivalent to accepting the DRMST null hypothesis $l^{\Theta}(\theta^*) = l^{\Psi}(\psi^*)$ under the assumption the Variance MST null hypothesis is false.

Theorem 5 in the Appendix shows that if the significance level used for both the Variance MST and the Direct Difference Loss MST in the testing procedure is α, then the Type 1 error probability for this two-stage testing procedure is asymptotically bounded by α as well. In addition, Theorem 5 shows that the Type 2 error probability for this two-stage testing procedure approaches zero as the sample size becomes large.

2. Assumptions

The following assumptions described in this section establish the following properties associated with the statistical tests developed in the previous section. Specifically, these assumptions are used in the Appendix of this article to demonstrate that as the sample size becomes sufficiently large: (a) the probability of a Type 1 error is asymptotically bounded from above by a pre-specified significance level, and (b) the probability of a Type 2 error converges to zero. These assumptions are also used in the Appendix to establish that the critical estimators in the analysis are strongly consistent estimators. It should be emphasized that these conclusions establish only minimal conditions for the legitimacy of these *large sample* testing procedures. However, considerable empirical work has provided strong support that many of these testing procedures are effective in specialized applications for typical sample sizes (e.g., see Barth, Beaver, & Landsman, 1998; Chen & Plott, 1998; Gertham, 1997; Golden, 1995, 1998; Grootendorst, 1995; Vincent, 1999; Wang, Halbrendt, & Johnson, 1996).

2.1. Data Generating Process Assumptions

Assumption A1. The sequence of d-dimensional real vectors defining the observed data $x_1, x_2, \ldots$ are a realization of the DGP, $\tilde{x}_1, \tilde{x}_2, \ldots$, on a complete probability space $(\mathcal{R}^{d\infty}, \mathcal{B}^{d\infty}, P_{DGP})$.

In practice, statistical tests must be computed using a data sample, X_n, consisting of n d-dimensional real vectors such that: $[x_1, \ldots, x_n] \subseteq \mathcal{R}^{dn}$. However, theoretical arguments based upon probability theory make claims only about the n d-dimensional random vectors $\tilde{x}_1, \ldots, \tilde{x}_n$ from the DGP. Thus, Assumption A1 provides the required theoretical linkage between the observed data (which is merely a sequence of n d-dimensional real vectors) and the underlying probability model associated with a stochastic sequence of n d-dimensional random vectors.

Additionally, Assumption A1 specifies constraints upon how the random vectors may be represented (i.e., the sample space modeling assumptions) through specification of the nature of the measurable space $(\mathcal{R}^{d\infty}, \mathcal{B}^{d\infty})$. This measurable space is assumed to be common to all probability models under consideration. The constraints on the measurable space are defined in a sufficiently general matter so that the probability models relevant to the theory described in this paper include probability models for continuous random variables, discrete random variables, and "mixed" continuous-discrete random variables. For further discussion of Assumption A1, please see White (1994, pp. 1–7).

Assumption A2. The DGP, $\tilde{x}_1, \tilde{x}_2, \ldots$, is bounded (i.e., there exists a finite constant K such that $|\tilde{x}_t| < K$ for all integer t w.p.1.).

The physically reasonable Assumption A2 implies that all observations generated by the DGP w.p.1 have a maximum magnitude. Thus, Assumption A2 is not applicable to stochastic processes which are not bounded (e.g., a sequence of i.i.d. Gaussian random variables). Assumption A2 is also immediately applicable to probability models such as logistic regression models which are defined with respect to finite sample spaces. Assumption A2 might appear to rule out inference involving probability models such as Gaussian models since such models do not have bounded sample spaces. In practice, however, this assumption is not restrictive (e.g., inference using Gaussian probability models *is* permissable) since the assumption that the probability models under consideration are correctly specified is not required. For example, define a DGP $\{\tilde{x}_t\}$ such that $\tilde{x}_t$ is the average of a large finite number of bounded uniformly distributed random variables with common finite mean and common finite positive variance. Then $\tilde{x}_t$ is bounded and has an approximate Gaussian distribution by the Central Limit Theorem. Such a DGP may be reasonably modeled by a Gaussian model.

Assumption A3. The DGP is τ-dependent which means that there exists a finite nonnegative integer τ such that $\tilde{\mathbf{x}}_s$ and $\tilde{\mathbf{x}}_t$ are independent for all $|s - t| > \tau$.

Assumption A3 corresponds to the physically reasonable assumption regarding the DGP which essentially states that two events separated by a sufficiently long time interval τ will be statistically independent. Note that the case of independent observations corresponds to the $\tau = 0$ case. Also note that the τ-dependent assumption rules out DGP stochastic processes defined implicitly by most stochastic dynamical systems. For example, consider the stochastic process $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots$ defined such that

$$\tilde{\mathbf{x}}_{t+1} = \mu \tilde{\mathbf{x}}_t + \tilde{\mathbf{n}}_t \tag{5}$$

where $\tilde{\mathbf{n}}_t$ is an i.i.d zero-mean bounded stochastic process. The stochastic process defined in (5) is *not* τ-dependent since $\tilde{\mathbf{x}}_t$ is always functionally dependent upon $\tilde{\mathbf{x}}_1$ as $t \to \infty$. On the other hand (if $|\mu| < 1$ in (5)) the functional dependence of $\tilde{\mathbf{x}}_t$ on $\tilde{\mathbf{x}}_1$ decreases as $t \to \infty$ suggesting that a probability model designed to represent a DGP of the form of (5) would yield a good approximation to a τ-dependent DGP.

Assumption A4. The DGP stochastic process, $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots$ is stationary (i.e., the joint distribution function of $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots$ is identical to the joint distribution function of $\tilde{\mathbf{x}}_{m+1}, \tilde{\mathbf{x}}_{m+2}, \ldots$ for $m = 1, 2, \ldots$).

Assumption A4 states that the joint distribution of the observed data is time-invariant. Although Assumption A4 implies that the observations are identically distributed, Assumption A4 does not require the observations to be independent. The combination of Assumption A4 with Assumption A3 defines a broad class of time-series DGP stochastic processes which includes the special case of i.i.d observations when $\tau = 0$.

Finally, it should be emphasized that using the methods of White (1994; also see Rivers and Vuong, 2002) that Assumptions A2 through A4 can be generalized to characterize a large class of asymptotically second-order stationary stochastic mixing processes with appropriate bounds on the moments of such processes. The relatively simple assumptions provided here, however, are substantially easier to understand and verify and yet are applicable to many important cases of interest. An important contribution of this paper is the presentation of the stronger assumptions embodied in A2 through A4 and an in-depth exploration of the interpretation and consequences of these strong assumptions.

2.2. Model Selection Criterion Assumptions

Assumption A5. Let $\Theta \subset \mathcal{R}^p$ be a convex, closed, bounded, and nonempty set. Assume the discrepancy loss $c^\Theta : \Theta \times \Gamma \to \mathcal{R}$ where $\Gamma \subseteq \mathcal{R}^d$.

Definition 1: A piecewise continuous function Let $\delta_0, \delta_1, \ldots, \delta_M$ be a finite set of $M + 1$ real finite numbers such that $\delta_{j-1} < \delta_j$, $j = 1, \ldots M$. Define

$$\mathcal{D}_j = \{\mathbf{x} = [x_1, \ldots, x_d] \in \mathcal{R}^d : \delta_{j-1} < x_i < \delta_j, i = 1, \ldots d\}$$

and $\bar{\mathcal{D}}_j$ as the closure of $\mathcal{D}_j$, $j = 1, \ldots M$. Let $\Gamma = \cup_{j=1}^M \bar{\mathcal{D}}_j$. Define a bounded function $f : \Gamma \to \mathcal{R}$ such that the restriction of f, f_j, on $\mathcal{D}_j$ is continuous ($j = 1, \ldots, M$). In addition, assume that each such f_j has the property that an extension of f_j to $\bar{\mathcal{D}}_j$ exists which is continuous on $\bar{\mathcal{D}}_j$, $j = 1, \ldots, M$. A function f with these properties is called a *piecewise continuous function* on Γ.

A vector-valued function $\mathbf{f} : \Gamma \to \mathcal{R}^d$ is said to be a piecewise continuous function on Γ if all d of its components are piecewise continuous functions on $\Gamma \subseteq \mathcal{R}^d$. Note that a continuous function on a closed subset of $\mathcal{R}^d$ is a piecewise continuous function on that closed subset as well.

An example application of the piecewise continuous function concept occurs in the context of considering probability models which possess "recoded" predictor variables. For example, consider a linear regression model $\tilde{y} = m\tilde{z} + b + \tilde{n}$ where $z = f(x)$ is the "recoded" predictor variable, $\tilde{y}$ is the dependent variable, $\tilde{n}$ is zero-mean Gaussian noise with known variance, and the parameter vector $\theta = (m, b)$. The function f is defined such that $f(x) = 1$ for $x > 0$, $f(x) = 0.5$ for $x = 0$, and $f(x) = 0$ for $x < 0$. The above definition of a piecewise continuous function can be used to verify that f is piecewise continuous. First, note that f is bounded. Now define $\mathcal{D}_1 = \{x : x < 0\}$ and $\mathcal{D}_2 = \{x : x > 0\}$. The restriction of f to $\mathcal{D}_1$, f_1, is continuous and the restriction of f to $\mathcal{D}_2$, f_2, is continuous. In addition, it is possible to extend f_1 to $\bar{\mathcal{D}}_1$ so that the extension of f_1, f_1^+, is continuous by choosing $f_1^+(0) = 0$. Similarly, the extension of f_2 to $\bar{\mathcal{D}}_2$, f_2^+, can be made continuous by choosing $f_2^+(0) = 1$.

Assumption A6. Let Γ be a closed subset of $\mathcal{R}^d$. Assume the discrepancy loss $c^\Theta : \Theta \times \Gamma \to \mathcal{R}$ has the property that $c^\Theta(\theta, \cdot)$, $\nabla c^\Theta(\theta, \cdot)$, and $\nabla^2 c^\Theta(\theta, \cdot)$ exist and are piecewise continuous on Γ for all $\theta \in \Theta$.

It can be shown that Assumption A6 (and Assumption A8 below) are actually stronger than necessary since all theorems and results in this paper are valid if one replaces the phrase "piecewise continuous function" with the phrase "Borel sigma-field measurable function." However, the stronger version of A6 (and A8 below) is used since: (a) every piecewise continuous function on a closed subset of a Euclidean vector space is a Borel sigma-field measurable function (see Theorem 2.5 of Gordon, 1994), (b) the piecewise continuous function concept is much more widely accessible to practitioners in the field, (c) from an engineering perspective little generality is lost by requiring that the relevant functions are piecewise continuous (instead of measurable), and (d) the stronger piecewise continuous assumption dramatically simplifies the other assumptions of this analysis as well.

Assumption A7. Let Γ be a closed subset of $\mathcal{R}^d$. Assume the discrepancy loss $c^\Theta : \Theta \times \Gamma \to \mathcal{R}$ has the property that for all $\mathbf{x} \in \Gamma$: $c^\Theta(\cdot, \mathbf{x})$, $\nabla c^\Theta(\cdot, \mathbf{x})$, and $\nabla^2 c^\Theta(\cdot, \mathbf{x})$ exist and are continuous on Θ.

Assumption A7 allows DRMST theory to be applicable to a broad range of differentiable discrepancy loss functions. It should be noted, however, that some important loss functions such as root mean square error loss functions (e.g., $c^\Theta(\theta, \mathbf{x}) = |\mathbf{x} - \theta|$, $\Theta = \{\theta : |\theta| < K, K > 0\}$) do not have a continuous well-defined gradient and therefore can not be handled within the DRMST theoretical framework.

Assumption A8. Let Γ^n be a closed, bounded, and convex subset of $\mathcal{R}^{dn}$. For each $n = 1, 2, \ldots$, assume $k_n^\Theta : \Theta \times \Gamma^n \to \mathcal{R}$ is a piecewise continuous function in both of its arguments.

Assumption A8 permits the *penalty term* $\tilde{k}_n^\Theta = k_n^\Theta(\tilde{\theta}_n, \tilde{\mathbf{X}}_n)$ to be functionally dependent upon the data sample $\tilde{\mathbf{X}}_n$ and/or the respective optimal parameter estimate $\tilde{\theta}_n$. The function k_n^Θ is called the *penalty function.*

Assumption A9. Assume that the penalty term $\tilde{k}_n^\Theta$ has the property that: $\sqrt{n}\tilde{k}_n^\Theta \to 0$ in probability as $n \to \infty$.

Assumptions A8 and A9 allow for the use of a large variety of model complexity penalty terms (see Sin and White, 1996, for a general review and general analysis) such as Akaike Infor-

mation Criterion (AIC) penalty terms (e.g., Akaike, 1973; Bozdogan, 1987; Linhart & Zucchini, 1986), Bayes Information Criterion (BIC) penalty terms (e.g., Djuric, 1998; Kass & Wasserman, 1995; Schwartz, 1978), and Minimum Descriptive Length (MDL) penalty terms (e.g., Balasubramanian, 1997; Clarke & Barron, 1990; Qian & Kunsch, 1998; Rissanen, 1996).

2.3. Solution Assumptions

Assumption A10. Assume for some θ^* in the interior of Θ that ∇l^{Θ} evaluated at θ^* is a vector of zeros. Assume $\nabla^2 l^{\Theta}$ is positive definite on Θ.

Assumption A10 is used to guarantee the existence and uniqueness of a solution $\omega^* = (\theta^*, \psi^*)$ and is relevant for constructing appropriate termination conditions for a given parameter estimation algorithm as described in section 1.1 (see Theorem 2 in the Appendix). Specifically, the assumption that the gradient of l^{Θ} vanishes at θ^* may be used as the basis for determining whether the parameter estimation algorithm has converged as discussed in section 1.1.2.

The assumption that the matrix-valued function $\nabla^2 l^{\Theta}$ is positive definite on Θ (and the corresponding assumption that $\nabla^2 l^{\Psi}$ is positive definite on Ψ) in Assumption $A10$ is used to guarantee that $\omega^* = (\theta^*, \psi^*)$ will be unique on Ω. Note that this assumption is formulated to facilitate the choice of $\Omega = \Theta \times \Psi$. In practice, when the discrepancy risk functions are convex it is convenient to choose Ω to be very large. When the discrepancy risk functions have multiple strict local minima, then Ω is typically chosen to be a sufficiently small neighborhood so that each model under consideration is associated with a unique strict local minimum (i.e., Θ and Ψ each contain a unique strict local minimum with respect to l^{Θ} and l^{Ψ} respectively).

The condition that $\nabla^2 l^{\Theta}$ is positive definite on Θ may be verified analytically in some situations by simply checking that $\theta^T[\nabla^2 l^{\Theta}]\theta$ is strictly positive for all $\theta \in \Theta$. If an analytical analysis is not possible, then $\nabla^2 l^{\Theta}(\theta^*)$ may be estimated by $\mathbf{A}_n^{\Theta}(\theta_n^*)$ as discussed in section 1.1.4. If $\mathbf{A}_n^{\Theta}(\theta_n^*)$ is approaching a positive definite matrix as the sample size n becomes large as discussed in section 1.1.4, then this suggests that θ^* is a strict local minimum. If θ^* is a strict local minimum, then this logically implies that a sufficiently small neighborhood, Θ, of θ^* exists such that θ^* is the unique global minimum on Θ.

Assumption A11. Assume $\mathbf{B}^*$ is positive definite.

The assumption that the matrix constant $\mathbf{B}^*$ is positive definite (see section 1.2 for the definition of $\mathbf{B}^*$) is required in order to apply an appropriate central limit theorem for dependent random variables.

Assumption A12. Assume $r^* \neq -1/(2\tau)$.

The quantity r^* in Assumption A12 is formally defined in section 1.3. Assumption A12 will usually be satisfied in practice. Also note, that if the observations are i.i.d. then Assumption A12 is satisfied (see sec. 1.3). If the observations are not i.i.d., then r^* can be estimated by r_n (see sec. 1.3) and empirically checked to see if $r_n \to 0$ w.p.1 as $n \to \infty$. An example of a situation where Assumption A12 is not satisfied occurs when $c^{\Theta}(\theta^*, \tilde{\mathbf{x}}_t) - c^{\Psi}(\psi^*, \tilde{\mathbf{x}}_t) = \tilde{\nu}_t - \tilde{\nu}_{t-1}$ where $\{\tilde{\nu}_t\}$ is a white noise process with strictly positive finite variance.[2]

3. Summary

This article has introduced an extension of Vuong's (1989; also see Golden, 2000) large sample model selection test theory called the Discrepancy Risk Model Selection Test (DRMST).

---

[2]The author gratefully acknowledges the contribution of this example from one of the referees for this paper.

The DRMST generalizes the quasi-maximum likelihood discrepancy measures considered by Vuong (1989) for the i.i.d. case to a substantially more general case which permits generalized discrepancy loss functions with penalty terms in the context of time-series data analysis. An important contribution of DRMST theory is that its assumptions are easily verifiable, yet the theory is broadly applicable to problems involving arbitrary twice continuously differentiable objective functions such as: log-likelihood type objective functions with model selection criteria such as AIC, BIC, and MDL.

DRMST theory is applicable to constructing large sample model selection tests for a number of important problems. First, a DRMST model selection test for comparing competing nested or nonnested probability models with different functional forms can be constructed. To illustrate this point, let $\{\tilde{n}_t\}$ denote a strictly stationary τ-dependent Gaussian stochastic process whose elements are identically distributed yet highly correlated. Using the DRMST theory, the linear regression model $\tilde{y}_t = m_1\tilde{x}_t + b_1 + \tilde{n}_t$ can be compared with the nonlinear regression model $\tilde{y}_t = (m_2\tilde{x}_t + b_2)^2 + \tilde{n}_t$ with respect to the DGP $\{(\tilde{y}_t, \tilde{x}_t)\}$. Note these two probability models are not only nonnested but also have distinct functional forms. Second, a DRMST model selection test for comparing competing preprocessing transformations of the DGP for a given probability model can be constructed. Thus, for example, the model $\tilde{y}_t = m_1[\tilde{x}_t]^2 + b_1 + \tilde{n}_t$ can be compared with the model $\tilde{y}_t = m_2[\tilde{x}_t]^4 + b_2 + \tilde{n}_t$ in order to determine which preprocessing transformation is more appropriate. Third, a DRMST model selection test for determining which of two strict local minima is "deeper" can be constructed as well through an appropriate choice of the joint model parameter space Ω. For example, this latter case would be relevant for log-likelihood loss functions associated with nonlinear regression models of the form: $\tilde{y}_t = \text{sine}(a\tilde{x}_t) + 2\cos(b\tilde{x}_t) + \tilde{n}_t$ $(\theta = [a, b])$ which have multiple strict local minima. And finally, all of the above analyses are applicable to misspecified probability models of a broad class of stationary stochastic processes. Although some of the calculations associated with DRMST theory require considerable programming effort, user-friendly software implementations should be readily available in the near future for linear, logistic, and multinomial logistic regression models (e.g., Martingale Research, 2001).

Appendix

Theorems

Let $\Gamma \subseteq \mathcal{R}^d$. Let $c^\Theta : \Theta \times \Gamma \to \mathcal{R}$ be some piecewise-continuous function on Θ and Γ. Then the *random function*, $\tilde{c}_i^\Theta$, corresponding to c^Θ maps an element, $\theta \in \Theta \subseteq \mathcal{R}^p$, into the random variable $c^\Theta(\theta, \tilde{x}_i)$.

Define $l^\Theta : \Theta \to \mathcal{R}$ such that for all $\theta \in \Theta$: $l^\Theta(\theta) = E\{c^\Theta(\theta, \tilde{x}_i)\}$ where the expectation is taken with respect to the DGP. Let $\tilde{l}_n^\Theta$ be a random function with the property that: $\tilde{l}_n^\Theta \to l^\Theta$ uniformly on Θ w.p.1 as $n \to \infty$, then the random function $\tilde{l}_n^\Theta$ is said to be a *strongly consistent* estimator of the deterministic function l^Θ.

Theorem 1: Estimator Consistency. Assume Assumptions A1 through A9 are satisfied for loss functions c^Θ and c^Ψ with respect to a particular DGP $\{\tilde{x}_t\}$. Then the random functions

$$\tilde{l}_n^\Theta, \tilde{l}_n^\Psi, \tilde{g}_n^\Theta, \tilde{g}_n^\Psi, \tilde{A}_n^\Theta, \tilde{A}_n^\Psi, \tilde{B}_n^\Theta, \tilde{B}_n^\Psi, \tilde{B}_n^{\Theta,\Psi}, \tilde{B}_n^{\Psi,\Theta}, \tilde{r}_n, \tilde{\sigma}_{Z_n}^2, \quad \text{and} \quad \tilde{R}_n$$

are strongly consistent estimators of the following list of continuous functions:

$$l^\Theta = E\{c^\Theta(\cdot, \tilde{x}_i)\}, l^\Psi = E\{c^\Psi(\cdot, \tilde{x}_i)\}, g^\Theta = E\{\nabla c^\Theta(\cdot, \tilde{x}_i)\},$$

$$g^\Psi = E\{\nabla c^\Psi(\cdot, \tilde{x}_i)\}, A^\Theta = E\{\nabla^2 c^\Theta(\cdot, \tilde{x}_i)\}, A^\Psi = E\{\nabla^2 c^\Psi(\cdot, \tilde{x}_i)\},$$

$$B^\Theta = \sum_{j=-\tau}^{\tau} E\{\nabla c^\Theta(\cdot, \tilde{x}_i)[\nabla c^\Theta(\cdot, \tilde{x}_{i+j})]^T\},$$

$$\mathbf{B}^{\Psi} = \sum_{j=-\tau}^{\tau} E\{\nabla c^{\Psi}(\cdot, \tilde{\mathbf{x}}_i)[\nabla c^{\Psi}(\cdot, \tilde{\mathbf{x}}_{i+j})]^T\},$$

$$\mathbf{B}^{\Theta,\Psi} = \sum_{j=-\tau}^{\tau} E\{\nabla c^{\Theta}(\cdot, \tilde{\mathbf{x}}_i)[\nabla c^{\Psi}(\cdot, \tilde{\mathbf{x}}_{i+j})]^T\},$$

$\mathbf{B}^{\Psi,\Theta} = [\mathbf{B}^{\Theta,\Psi}]^T$, r^* as defined in section 1.3,

$$\sigma_Z^2 = E\{(\tilde{c}_i^{\Theta} - \tilde{c}_i^{\Psi})^2\} + 2\sum_{j=1}^{\tau} E\{(\tilde{c}_i^{\Theta} - \tilde{c}_i^{\Psi})(\tilde{c}_{i+j}^{\Theta} - \tilde{c}_{i+j}^{\Psi})\}, \tag{6}$$

and

$$\mathbf{R} = \begin{bmatrix} -\mathbf{B}^{\Theta}[\mathbf{A}^{\Theta}]^{-1} & -\mathbf{B}^{\Theta,\Psi}[\mathbf{A}^{\Psi}]^{-1} \\ \mathbf{B}^{\Psi,\Theta}[\mathbf{A}^{\Theta}]^{-1} & \mathbf{B}^{\Psi}[\mathbf{A}^{\Psi}]^{-1} \end{bmatrix}$$

respectively.

Proof of Theorem 1. The proof of Theorem 1 follows immediately from using Assumptions A1 through A9, the definitions of White (1984, pp. 40–41), and Theorem A.2.2 from White (1994, pp. 351–352). ☐

Theorem 2: Consistency of Estimates. Assume Assumptions A1 through A10 are satisfied for loss functions c^{Θ} and c^{Ψ} and penalty terms $\tilde{k}_n^{\Theta}$ and $\tilde{k}_n^{\Psi}$ with respect to a particular DGP $\{\tilde{\mathbf{x}}_t\}$. Define $\tilde{\theta}_n^* = \operatorname{argmin} \tilde{l}_n^{\Theta}$ on Θ. Define $\tilde{\psi}_n^* = \operatorname{argmin} \tilde{l}_n^{\Psi}$ on Ψ. Then $\tilde{\omega}_n^* = (\tilde{\theta}_n^*, \tilde{\psi}_n^*) \to \omega^*$ w.p.1.

Proof of Theorem 2. Using Theorem 1 (this paper) and Theorem 3.4 of White (1994, p. 28), and Assumptions A1 through A10 it follows that $\tilde{\omega}_n^* \to \omega^*$ with probability one as $n \to \infty$ since ω^* is the unique global minimum on Ω by Assumption A10. ☐

Theorem 3: Variance MST. Assume Assumptions A1 through A12 are satisfied for loss functions c^{Θ} and c^{Ψ} and penalty terms $\tilde{k}_n^{\Theta}$ and $\tilde{k}_n^{\Psi}$ with respect to a particular DGP $\{\tilde{\mathbf{x}}_t\}$. Define $n\tilde{\sigma}_{Z_n}^2(\tilde{\omega}_n^*, \tilde{\mathbf{X}}_n)$ with respect to c^{Θ}, c^{Ψ}, $\tilde{k}_n^{\Theta}$, $\tilde{k}_n^{\Psi}$, and the random sample $\tilde{\mathbf{X}}_n$ from the DGP as described in section 1.4. Let $\tilde{w}_{i,n}^*$ be the i-th squared eigenvalue of the $p+q$-dimensional random matrix $\tilde{\mathbf{R}}_n^*$ defined in section 1.4 (i.e., the function $\tilde{\mathbf{R}}_n^*$ evaluated at $\tilde{\omega}_n^*$). Then $\tilde{w}_{i,n}^*$ converges to the i-th element, w_i^*, of a $p+q$-dimensional vector, $\mathbf{w}^*$, of strictly positive real numbers w.p.1 as $n \to \infty$ $(i = 1, \ldots, (p+q))$. If the Variance MST null hypothesis holds, then $n\tilde{\sigma}_{Z_n}^2(\tilde{\omega}_n^*, \tilde{\mathbf{X}}_n)$ has an asymptotic weighted chi-square distribution with weighting vector $\mathbf{w}^*$. If the Variance MST null hypothesis is false, then $n\tilde{\sigma}_{Z_n}^2(\tilde{\omega}_n^*, \tilde{\mathbf{X}}_n) \to \infty$ w.p.1 as $n \to \infty$.

Proof of Theorem 3. Please see following sections of this Appendix. ☐

Theorem 4: Direct Difference Loss MST. Assume Assumptions A1 through A12 are satisfied for loss functions c^{Θ} and c^{Ψ} and penalty terms $\tilde{k}_n^{\Theta}$ and $\tilde{k}_n^{\Psi}$ with respect to a particular DGP $\{\tilde{\mathbf{x}}_t\}$. Define $\tilde{Z}_n$ with respect to c^{Θ}, c^{Ψ}, $\tilde{k}_n^{\Theta}$, $\tilde{k}_n^{\Psi}$, and the random sample $\tilde{\mathbf{X}}_n$ from the DGP as described in section 1.4 such that Z_n in section 1.5 is a particular realization of $\tilde{Z}_n$. Assume the Variance MST Null Hypothesis is false. Then, given the null hypothesis of the Direct Difference Loss MST is true, $\sqrt{n}\tilde{Z}_n$ converges in distribution to a Gaussian random variable with mean zero and variance one as $n \to \infty$. If the null hypothesis of the Direct Difference Loss MST is false, then $\sqrt{n}\tilde{Z}_n \to \infty$ w.p.1 as $n \to \infty$.

Proof of Theorem 4. Define $\Delta_n^* = C_n^\Theta - C_n^\Psi$. Define $\tilde{d}_i^* = c^\Theta(\theta^*, \tilde{\mathbf{x}}_i) - c^\Psi(\psi^*, \tilde{\mathbf{x}}_i)$. Let, $\mathbf{g}_n^-$, denote the gradient of Δ_n^* implying that:

$$\mathbf{g}_n^-(\omega^*, \tilde{\mathbf{X}}_n) = \left[\nabla l_n^\Theta(\theta^*, \tilde{\mathbf{X}}_n), \quad -\nabla l_n^\Psi(\psi^*, \tilde{\mathbf{X}}_n)\right].$$

Then, using the mean-value theorem,

$$\Delta_n(\tilde{\omega}_n^*, \tilde{\mathbf{X}}_n) = \left(\frac{1}{n}\right)\sum_{i=1}^n \left[\tilde{d}_i^*\right] + \mathbf{g}_n^-(\omega^*, \tilde{\mathbf{X}}_n)^T[\tilde{\omega}_n^* - \omega^*] + \tilde{\epsilon}_n \tag{7}$$

where $\tilde{\epsilon}_n = o_p(1)O(|\tilde{\omega}_n^* - \omega^*|)$ is the remainder term and the notation $o_p(1)$ denotes a term converging in probability to zero by Theorem 2. Note that by Lemma 1 in the Appendix, $\sqrt{n}[\tilde{\omega}_n^* - \omega^*]$ converges in distribution and so is bounded in probability this implies that $\sqrt{n}\tilde{\epsilon}_n \to 0$ in probability as $n \to \infty$. $\qquad\square$

By Theorem 1 and Assumption A10, it follows that $\mathbf{g}_n^-(\omega^*, \tilde{\mathbf{X}}_n) \to \mathbf{0}$ with probability one as $n \to \infty$. Using this observation with (7) gives

$$\sqrt{n}\tilde{\Delta}_n = \sqrt{n}\Delta_n(\tilde{\omega}_n^*, \tilde{\mathbf{X}}_n) = \sqrt{n}\left(\frac{1}{n}\right)\sum_{i=1}^n \tilde{d}_i^* + \tilde{e}_n$$

where $\tilde{e}_n = O(\sqrt{n}|\tilde{\omega}_n^* - \omega^*|)o_p(1)$ where $o_p(1)$ denotes a term converging in probability to zero. Thus, $\tilde{e}_n$ converges in probability to zero since $O(|\sqrt{n}|\tilde{\omega}_n^* - \omega^*|)$ is bounded in probability.

Since the Variance MST null hypothesis

$$H_0^{\sigma^2}(\tau) : c^\Theta(\theta^*, \tilde{\mathbf{x}}_t) = c^\Psi(\psi^*, \tilde{\mathbf{x}}_t)$$

w.p.1 is false by assumption, it follows that the variance of $\Delta_n(\omega^*, \tilde{\mathbf{X}}_n)$, $\sigma_Z^2(\omega^*, \tilde{\mathbf{X}}_n)$, is strictly positive given the null hypothesis $H_0 : \Delta = l^\Theta(\theta^*) - l^\Psi(\psi^*) = 0$ holds. Since σ_Z^2 is a continuous function on the compact set Ω, σ_Z^2 is also a bounded function on Ω as well.

Using Assumptions A1 through A4 the result that $\tilde{e}_n = o_p(1)$, White's (1984; Theorem 5.19, p. 124) central limit theorem for dependent variables, and Slutsky's theorem (Serfling, 1980, p. 19) it follows that $\sqrt{n}\tilde{\Delta}_n$ has a mean zero asymptotic normal distribution with finite positive variance σ_Z^2 under H_0. If H_0 is false, then $\tilde{\Delta}_n$ converges to a finite constant with probability one by Theorem 1 implying that $\sqrt{n}\tilde{\Delta}_n \to \infty$ with probability one as $n \to \infty$.

Theorem 1. (DRMST) Assume Assumptions A1 through A12 are satisfied for loss functions c^Θ and c^Ψ and penalty terms $\tilde{k}_n^\Theta$ and $\tilde{k}_n^\Psi$ with respect to a particular DGP $\{\tilde{\mathbf{x}}_t\}$. Define the null hypothesis $H_0 : l^\Theta(\theta^*) = l^\Psi(\psi^*)$. Assume the DRMST is used to test H_0. Then, as $n \to \infty$, $p_n(\text{accept } H_0 \mid H_0 \text{ false}) \to 0$. In addition, $p_n(\text{reject } H_0 \mid H_0 \text{ true}) < \alpha$ as $n \to \infty$.

Proof of Theorem 5. The proof of Theorem 5 follows arguments presented in Vuong (1989). Let the notation $p_n[\text{VAR}_1]$ and $p_n[\text{VAR}_2]$ denote the estimated probability of Type 1 and Type 2 errors respectively for the Variance MST based upon a particular sample size n. Let the notation $p_n[\text{DDL}_1]$ and $p_n[\text{DDL}_2]$ denote the estimated probability of Type 1 and Type 2 errors respectively for the Direct Difference Loss MST based upon a particular sample size n. Let the notation $p_n[\text{DRMST}_1]$ and $p_n[\text{DRMST}_2]$ denote the estimated probability of Type 1 and Type 2 errors respectively for the DRMST based upon a particular sample size n.

By Theorem 3, $p_n[\text{VAR}_1] < \alpha$ as $n \to \infty$. By Theorem 4,

$$p_n[\text{DDL}_1] < \alpha \quad \text{as} \quad n \to \infty.$$

From the definition of the DRMST (see Figure 2), the DRMST procedure rejects H_0 if and only if both the Direct Difference Loss MST and Variance Test hypotheses are rejected. This implies that

$$p_n[\text{DRMST}_1] \leq \max\{p_n[\text{VAR}_1], p_n[\text{DDL}_1]\}.$$

Thus, it follows that $p_n[\text{DRMST}_1] < \alpha$ as $n \to \infty$ since $p_n[\text{VAR}_1]$ and $p_n[\text{DDL}_1]$ are asymptotically bounded by the significance level α.

The Type 2 error probability for the DRMST is now considered. Let the Variance MST null hypothesis be denoted by $H_0^{\sigma^2}$. Let $\neg H_0$ denote the assertion that the null hypothesis H_0 is false. From the definition of the DRMST (see Figure 2),

$$p_n[\text{DRMST}_2] = p_n[\text{Accept } H_0^{\sigma^2} \text{ OR (Reject } H_0^{\sigma^2} \text{ AND Accept } H_0) \mid \neg H_0]$$

$$p_n[\text{DRMST}_2] \leq p_n[\text{Accept } H_0^{\sigma^2} \mid \neg H_0] + p_n[\text{Reject } H_0^{\sigma^2} \text{ AND Accept } H_0 \mid \neg H_0]$$

Note that since $\neg H_0 \subseteq \neg H_0^{\sigma^2}$,

$$p_n[\text{Accept } H_0^{\sigma^2} \mid \neg H_0] \leq p_n[\text{Accept } H_0^{\sigma^2} \mid \neg H_0^{\sigma^2}] = p_n[\text{VAR}_2].$$

Note that

$$p_n[\text{Reject } H_0^{\sigma^2} \text{ AND Accept } H_0 \mid \neg H_0] = p_n[\text{Reject } H_0^{\sigma^2} \mid \neg H_0] p_n[\text{DDL}_2]$$

since

$$p_n[\text{DDL}_2] = p_n[\text{Accept } H_0 \mid \neg H_0 \text{ AND Reject } H_0^{\sigma^2}].$$

Thus,

$$p_n[\text{DRMST}_2] \leq p_n[\text{VAR}_2] + p_n[\text{DDL}_2] \tag{8}$$

The first term on the right-hand side of (8) converges to zero as $n \to \infty$ by Theorem 3. The second term on the right-hand side of (8) converges to zero as $n \to \infty$ by Theorem 4. $\qquad \square$

Lemmas

Lemma 1: Asymptotic Normality of Parameter Estimates. Given Assumptions A1 through A11,

$$\sqrt{n}[\tilde{\mathbf{B}}_n^*]^{-1/2}\tilde{\mathbf{A}}_n^*[\tilde{\omega}_n^* - \omega^*] \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Proof of Lemma 1. Let $\tilde{\mathbf{g}}_n(\omega^*) = \mathbf{g}_n(\omega^*, \tilde{\mathbf{X}}_n)$ be defined as

$$\left[\nabla l_n^{\Theta}(\theta^*, \tilde{\mathbf{x}}_n), \nabla l_n^{\Psi}(\psi^*, \tilde{\mathbf{x}}_n)\right].$$

Note that $E[\mathbf{g}_n(\omega^*, \tilde{\mathbf{X}}_n)] = \mathbf{0}$ by Assumption A10. Thus, $\mathbf{B}^*$ is the asymptotic covariance of $\sqrt{n}\mathbf{g}_n(\omega^*, \tilde{\mathbf{X}}_n)$. Note that $\mathbf{B}^*$ is finite since $\mathbf{B}$ is a continuous function on the compact set Ω. Let $\mathbf{h}$ be some real vector of the same dimension as $\mathbf{B}^*$ such that $\mathbf{h}^T\mathbf{h} = 1$. Then from the results of Theorem 1 and the central limit theorem for dependent variables (e.g., Theorem 5.19 in White, 1984, p. 124; also see Serfling, 1968), it follows that $\sqrt{n}\mathbf{h}^T[\mathbf{B}^*]^{-1}\tilde{\mathbf{g}}_n(\omega^*)$ converges in distribution to a random variable which may be represented as $\mathbf{h}^T\tilde{\mathbf{g}}^*$ where $\tilde{\mathbf{g}}^*$ is a multivariate Gaussian random vector with mean $\mathbf{0}$ and identity covariance matrix $\mathbf{I}$. Using the Cramer-Wold device (e.g., Proposition 5.1; White, 1984, p. 108), it then immediately follows that $\sqrt{n}[\mathbf{B}^*]^{-1/2}\tilde{\mathbf{g}}_n(\omega^*) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Also from Theorem 1, $\tilde{\mathbf{A}}_n^{\Theta}$ is a strongly consistent estimator of $\mathbf{A}$, and Assumption A10 implies

that $\mathbf{A}^*$ is positive definite. The Lemma's conclusion then immediately follows from White's (1994, pp. 89–90) Theorem 6.2. $\qquad\square$

Lemma 2: Vuong, 1989. Define

$$\mathbf{U} = \begin{bmatrix} \mathbf{B}^{\Theta} & -\mathbf{B}^{\Theta,\Psi} \\ -\mathbf{B}^{\Psi,\Theta} & \mathbf{B}^{\Psi} \end{bmatrix}$$

and

$$\mathbf{Q} = \begin{bmatrix} \mathbf{A}^{\Theta} & \mathbf{0}_{q,p} \\ \mathbf{0}_{p,q} & -\mathbf{A}^{\Psi} \end{bmatrix}.$$

Let $\mathbf{U}^*$ be defined as $\mathbf{U}$ evaluated at ω^*. Define $\mathbf{C}^* = [\mathbf{A}^*]^{-1}\mathbf{B}^*[\mathbf{A}^*]^{-1}$. Let $\mathbf{Q}^*$ be $\mathbf{Q}$ evaluated at ω^*. Then, $\mathbf{U}^*\mathbf{C}^* = (\mathbf{R}^*)^2$ and $\mathbf{U}^* = \mathbf{Q}^*\mathbf{C}^*\mathbf{Q}^*$.

Proof of Lemma 2 (Vuong, 1989). Note $\mathbf{Q}^*\mathbf{C}^* = -\mathbf{R}^*$, $\mathbf{U}^* = \mathbf{Q}^*\mathbf{C}^*\mathbf{Q}^*$, and therefore $\mathbf{U}^*\mathbf{C}^* = \mathbf{Q}^*\mathbf{C}^*\mathbf{Q}^*\mathbf{C}^* = (-\mathbf{R}^*)^2$. $\qquad\square$

Lemma 3. Assume Assumptions A1 through A6, A12 holds. The null hypothesis for the Variance MST $(\sigma_Z^2(\omega^*) = 0)$ is equivalent to the null hypothesis that $c^{\Theta}(\theta^*, \tilde{\mathbf{x}}_t) = c^{\Psi}(\psi^*, \tilde{\mathbf{x}}_t)$ w.p.1.

Proof of Lemma 3. Assumptions A1 through A6 are required to guarantee the existence of the expectations in Assumption (6). The null hypothesis for the Variance MST is given by $\sigma_Z^2(\omega^*) = 0$ where σ_Z^2 is defined in (6). Using the definition of r^* in section 1.3, it follows after some algebra that $\sigma_Z^2(\omega^*) = 0$ is satisfied if and only if either: (1) $r^* = -1/(2\tau)$, or (2) $c^{\Theta}(\theta^*, \tilde{\mathbf{x}}_t) = c^{\Psi}(\psi^*, \tilde{\mathbf{x}}_t)$ w.p.1. Assumption A12 $(r^* \neq -1/(2\tau))$ then rules out the possibility that $r^* = -1/(2\tau)$. $\qquad\square$

Lemma 4. If Assumptions A1 through A11 are satisfied, then as $n \to \infty$, the i-th element of the $p + q$-dimensional matrix $\tilde{\mathbf{R}}_n^*$, $\tilde{w}_{i,n}^*$, converges w.p.1 to the i-th element of $\mathbf{R}^*$, w_i^*. In addition, w_i^*, is strictly positive for $i = 1, \ldots, (p+q)$.

Proof of Lemma 4. Since the eigenvalues of the $\mathbf{R}$ matrix are continuous functions of the matrix's elements (Franklin, 1968, pp. 191–192), it follows that the i-th squared eigenvalue of the $p + q$-dimensional matrix $\tilde{\mathbf{R}}_n^*$, $\tilde{w}_{i,n}^*$, converges w.p.1 to the i-th squared eigenvalue of $\mathbf{R}^*$, w_i^*. It is now necessary to show that w_i^* is strictly positive for $i = 1, \ldots, (p+q)$. By Lemma 2, $(\mathbf{Q}^*\mathbf{C}^*)^2 = (\mathbf{R}^*)^2$. The matrix $\mathbf{C}^*$ is strictly positive definite since $\mathbf{A}^*$ and $\mathbf{B}^*$ are strictly positive definite by Assumptions A10 and A11 and the definition of $\mathbf{C}^*$. The matrix $\mathbf{Q}^*\mathbf{C}^*$ is a full rank real square matrix since $\mathbf{Q}^*$ is full rank (since $\mathbf{A}^*$ is positive definite and the definition of $\mathbf{Q}^*$). Thus, every eigenvalue of $\mathbf{Q}^*\mathbf{C}^*$ is either strictly positive or strictly negative. This implies that all eigenvalues of $(\mathbf{Q}^*\mathbf{C}^*)^2 = (\mathbf{R}^*)^2$ are strictly positive. $\qquad\square$

Calculations of Variance Function Gradient and Hessian

Derivation of the variance function gradient. The gradient of $\sigma_{Z_n}^2$ on Ω, $\nabla\sigma_{Z_n}^2 : \Omega \times \Gamma^n \to \mathcal{R}^{p+q}$ is given by:

$$\nabla\sigma_{Z_n}^2 = \frac{2}{n}\sum_{i=1}^{n}\sum_{j \in J_{i,\tau}} \begin{bmatrix} (c_j^{\Theta} - c_j^{\Psi})\nabla c_i^{\Theta} \\ (c_j^{\Psi} - c_j^{\Theta})\nabla c_i^{\Psi} \end{bmatrix}.$$

Derivation of the variance function Hessian. The Hessian of $\sigma_{Z_n}^2$ on Ω, $\nabla^2 \sigma_{Z_n}^2 : \Omega \times \Gamma^n \rightarrow$ $\mathcal{R}^{(p+q)\times(p+q)}$: is given by $\nabla^2 \sigma_{Z_n}^2 = 2\mathbf{U}_n + \mathbf{E}_n$ where

$$\mathbf{U}_n = \begin{bmatrix} \mathbf{B}_n^\Theta & -\mathbf{B}_n^{\Theta,\Psi} \\ -\mathbf{B}_n^{\Psi,\Theta} & \mathbf{B}_n^\Psi \end{bmatrix}$$

and

$$\mathbf{E}_n = \frac{2}{n} \sum_{i=1}^n \sum_{j \in J_{i,\tau}} \begin{bmatrix} (c_j^\Theta - c_j^\Psi)\nabla^2 c_i^\Theta & \mathbf{0}_{p,q} \\ \mathbf{0}_{q,p} & (c_j^\Psi - c_j^\Theta)\nabla^2 c_i^\Psi \end{bmatrix}.$$

Proof of Theorem 3

By Lemma 3, it follows that if the null hypothesis of the Variance MST is false, then $\sigma_Z^2(\omega^*) > 0$. Thus, $n\tilde{\sigma}_{Z_n}^2(\tilde{\omega}_n^*) \rightarrow \infty$ w.p.1 as $n \rightarrow \infty$ by Theorem 1 and Theorem 2. This proves the theorem for the case where the Variance MST null hypothesis is false. Thus, in the remainder of the proof, it is assumed that the Variance MST null hypothesis is true.

Using $\nabla^2 \sigma_{Z_n}^2 = 2\mathbf{U}_n + \mathbf{E}_n$, H_0, and the definition of $\mathbf{U}$ from Lemma 2 we have,

$$\nabla^2 \tilde{\sigma}_{Z_n}^2 = 2\mathbf{U}^* + o_p(1) \tag{9}$$

evaluated at $\tilde{\omega}_n^*$ where $\mathbf{U}^*$ is $\mathbf{U}$ evaluated at ω^*.

Using the mean-value theorem and (9) and $\sigma_n^* = \sigma_{Z_n}(\omega^*, \tilde{\mathbf{X}}_n)$,

$$\sigma_{Z_n}^2(\tilde{\omega}_n^*, \tilde{\mathbf{X}}_n) = [\sigma_n^*]^2 + \nabla \sigma_{Z_n}^2(\omega^*, \tilde{\mathbf{X}}_n)(\tilde{\omega}_n^* - \omega^*) + [\tilde{\omega}_n^* - \omega^*]^T \mathbf{U}^*[\tilde{\omega}_n^* - \omega^*] + \tilde{\epsilon}_n \tag{10}$$

where $\tilde{\epsilon}_n = o_p(1)O(|\tilde{\omega}_n^* - \omega^*|^2)$ is the remainder term.

Using Lemma 3 and the assumption that the Variance MST null hypothesis is true, the first two terms in (10) are equal to zero with probability one for all integer n. Thus, multiplying by n:

$$n\tilde{\sigma}_{Z_n}^2(\tilde{\omega}_n^*, \tilde{\mathbf{X}}_n) = n[\tilde{\omega}_n^* - \omega^*]^T \mathbf{U}^*[\tilde{\omega}_n^* - \omega^*] + \tilde{e}_n \tag{11}$$

where $\tilde{e}_n = o_p(1)O(n|\tilde{\omega}_n^* - \omega^*|^2)$ is a remainder term which converges in probability to zero since $n|\tilde{\omega}_n^* - \omega^*|^2$ converges in distribution by Lemma 2 and thus is bounded in probability.

Also note that by Lemma 1, $\sqrt{n}(\tilde{\mathbf{C}}_n^*)^{-1/2}(\tilde{\omega}_n^* - \omega^*) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Let w_i^* denote the i-th eigenvalue of $(\mathbf{R}^*)^2 = \mathbf{U}^*\mathbf{C}^*$, $i = 1, \ldots, (p+q)$. Let $\tilde{z}_i$ denote a zero-mean, unit-variance Gaussian random variable. Thus, it follows from Lemma 1, Vuong's (1989) Lemma 3.2, and Slutsky's Theorem that:

$$n\tilde{\sigma}_{Z_n}^2(\tilde{\omega}_n^*, \tilde{\mathbf{X}}_n) = \sum_{i=1}^{p+q} w_i^*(\tilde{z}_i)^2 + o_p(1)O(n|\tilde{\omega}_n^* - \omega^*|^2).$$

The result of the theorem then directly follows by using Lemma 1, noting that $\tilde{w}_{i,n}^* \rightarrow w_i^*$ with probability one (Lemma 4) and then applying Slutsky's Theorem again.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrox & F. Caski (Eds.), *Second International Symposium on Information Theory* (p. 267). Budapest: Akademiai Kiado.

Balasubramian, V. (1997). Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation, 9*, 349–368.

Barth, M.E., Beaver, W.H., Landsman, W.R. (1998). Relative valuation roles of equity book value and net income as a function of financial health. *Journal of Accounting & Economics, 25*, 1–34.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345–370.

Chen, K.Y., & Plott, C.R. (1998). Nonlinear behavior in sealed bid first price auctions. *Game Theory and Economic Behavior, 25*, 34–78.

Clarke, S., & Barron, A.R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory, IEEE-IT, 36*, 453–471.

Davies, R.B. (1980). The distribution of a linear combination of χ^2 random variables. *Applied Statistics, 29*, 323–333.

Djuric, P.M. (1998). Asymptotic MAP criteria for model selection. *IEEE Transactions on Signal Processing, 46*, 2726–2735.

Draper, N.R., & Smith, H. (1981). *Applied regression analysis*. New York, NY: Wiley.

Efron, B. (1984). Comparing nonnested linear models. *Journal of the American Statistical Association, 79*, 791–803.

Franklin, J.N. (1968). *Matrix theory*. New Jersey, NJ: Prentice-Hall.

Foutz, R.V., & Srivastava, R.C. (1977). The performance of the likelihood ratio test when the model is incorrect. *Annals of Statistics, 5*, 1183–1194.

Gan, L., & Jiang, J. (1999). A test for global maximum. *Journal of the American Statistical Association, 94*, 847–855.

Gertham, U.G. (1997). Equity in health care utilization: Further tests based on hurdle models and Swedish micro data. *Health Economics, 6*, 303–319.

Golden, R.M. (1995). Making correct statistical inferences using a wrong probability model. *Journal of Mathematical Psychology, 38*, 3–20.

Golden, R.M. (1996). *Mathematical methods for neural network analysis and design*. Cambridge, MA: MIT Press.

Golden, R.M. (1998). Knowledge digraph contribution analysis of protocol data. *Discourse Processes, 25*, 179-210.

Golden, R.M. (2000). Statistical tests for comparing possibly misspecified and nonnested models. *Journal of Mathematical Psychology, 44*, 153–170.

Gordon, R.A. (1994). *The integrals of Lebesgue, Denjoy, Perron, and Henstock*. Providence, RI: American Mathematical Society.

Grootendorst, P.V. (1995). A comparison of alternative models of prescription drug utilization. *Health Economics, 4*, 183–198.

Kass, R.E. & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association, 90*, 928–934.

Larson, H.J., & Shubert, B.O. (1979). *Probabilistic models in engineering sciences: Volume 1*. New York, NY: Wiley.

Linhart, H. (1988). A test whether two AIC's differ significantly. *South African Statistical Journal, 22*, 153–161.

Linhart, H., & Zucchini, W. (1986). *Model selection*. New York, NY: Wiley.

Martingale Research. (1997). *CCR Data Modeling* [Computer software]. Bedford, TX: Author. (Information about other DRMST regression modeling software is available at http://www.martingale-research.com).

Myung I., Forster M., & Browne M. (2000). Special issue on model selection. *Journal of Mathematical Psychology, 44*, 1–2.

Qian, G. & Kunsch, H.R. (1998). Some notes on Rissanen's stochastic complexity. *IEEE Transactions on Information Theory, 44*, 782–786.

Rissanen, J.J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, 42*, 40–47.

Rivers, D., & Vuong, Q. (2002). Model selection tests for nonlinear dynamic models. *The Econometrics Journal, 5*, 1-39.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Serfling, R.J. (1980). *Approximation theorems of mathematical statistics*. New York, NY: Wiley.

Sheil, J., & O'Muircheartaigh, I. (1977). The distribution of nonnegative quadratic forms in normal variables. *Applied Statistics, 26*, 92–98.

Shimodaira, H. (1997). Assessing the error probability of the model selection test. *Annals of the Institute of Statistical Mathematics, 49*, 395–410.

Sin, C., & White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics, 71*, 207–225.

Vincent, L. (1999). The information content of funds from operations (FFO) for real estate investment trusts (REITs). *Journal of Accounting & Economics, 26*, 69–104.

Vuong, Q.H. (1989). Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica, 57*, 307–333.

Vuong, Q.H., & Wang, W. (1993). Minimum chi-square estimation and tests for model selection. *Journal of Econometrics, 56*, 141–168.

Wang, Q.B., Halbrendt, C., & Johnson, S.R. (1996). A nonnested test of the AIDS vs. the translog demand system. *Economics Letters, 51*, 139–143.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*, 1–25.

White, H. (1984). *Asymptotic theory for econometricians*. San Diego, CA: Academic Press.

White, H. (1989). Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association, 84*, 1003–1013.

White, H. (1994). *Estimation, inference, and specification analysis*. New York, NY: Cambridge University Press.

Wilks, S.S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics, 9*, 60–62.