Statistical Tests for Comparing Possibly Misspecified and Non-nested Models

Richard M. Golden

University of Texas at Dallas

Send correspondence to:
Dr. Richard M. Golden
School of Human Development
University of Texas at Dallas
GR 41, Box 830688
Richardson, Texas 75083-0688
(972) 883-2423
Email: GOLDEN@UTDALLAS.EDU

Running head: Statistical Test for Model Selection

Abstract

A Model Selection Criteria (MSC) involves selecting the model with the best "estimated goodness-of-fit" to the data generating process. Following the method of Vuong (1989, *Econometrica, 57*, 307-333), a large sample Model Selection Statistical Test (MST) is introduced that can be used in conjunction with most existing MSC procedures to decide if the estimated goodness-of-fit for one model is significantly different from the estimated goodness-of-fit for another model. The MST extends the classical generalized likelihood ratio test, is valid in the presence of model misspecification, and is applicable to situations involving non-nested probability models. Simulation studies designed to illustrate the concept of the MST and its conservative decision rule (relative to the MSC method) are also presented.

Statistical Tests for Comparing Possibly Misspecified and Non-nested Models

An important problem in model selection is concerned with identifying the best-fitting model to some unobservable data generating process given only a data sample from that process (for further discussion see Akaike, 1973; Cox, 1962; Schwarz, 1978; Bozdogan, 1987; Linhart and Zucchini, 1986; Sin and White, 1996). In such procedures, the fit of each model to the data generating distribution is evaluated using some goodness-of-fit function. The model with the best goodness-of-fit is then selected. Such procedures permit multiple models to be simultaneously compared so that if there are K possible models, then there are K possible decisions: *select model 1*, *select model 2*, ..., *select model K*. Such procedures will be referred to as Model Selection Criteria (MSC) procedures.

Model Selection Test (MST) methodology. In some situations, however, it might be desirable to make the additional decision that *there is not sufficient evidence for selecting one model over another*. A Model Selection Test (MST) procedure is a statistical test specifically designed to test the null hypothesis that all probability models fit the data generating process equally well. Thus, a MST results in *K+1* possible decisions for a comparison among K models at a chosen significance level: *select model 1*, *select model 2*, ..., *select model K*, or *conclude there is not sufficient evidence for selecting one model over another.*

Comparing large numbers of models simultaneously. The MST procedures reviewed in this article are limited to comparing only two models at a time, whereas MSC procedures are not limited in this way. Of course, a MST for comparing multiple models simultaneously can be developed using multiple pair-wise MSTs but then the experiment-wise error rate will be inflated (see Golden, 1995, for further discussion of this point). In such cases where large numbers of multiple probability models must be simultaneously compared, a MSC methodology is usually preferable to the pair-wise MST methodology discussed in this article.

MST procedures versus "goodness-of-fit" tests. The MST procedures reviewed in this article are different in an important way from goodness-of-fit statis-

tical tests. A goodness-of-fit statistical test is concerned with testing the idealistic null hypothesis that a given probability model fits the underlying data generating process effectively. In contrast, the MST procedures reviewed here are concerned with testing the more pragmatic null hypothesis that two given probability models provide equally effective descriptions of the underlying data generating process even in situations where *neither* probability model is truly appropriate.

Historical comments. An important early approach to the MST problem was Wilk's (1938) Generalized Likelihood Ratio Test (GLRT) which tested the null hypothesis that two "fully nested" models were equivalent. More recently, Efron (1984) considered the problem of comparing two non-nested linear models. Linhart (1988) proposed a large sample statistical test for comparing "non-nested" models. Shimodaira (1997) emphasized the difficulty of applying Linhart's (1988) methodology to more general situations where models could be nested or partially nested, and proposed a modification of Linhart's test statistic to solve this problem.

An alternative (but closely related) approach to the method of Shimodaira (1997), was described earlier by Vuong (1989). Vuong's (1989) approach was largely influenced by the work of White (1982) who was concerned with the problem of making statistical inferences in the presence of model misspecification (see Golden, 1995, and White, 1994, for relevant reviews). Vuong's (1989) theory essentially combined the GLRT method and Linhart's (1988) method to obtain a two-stage large sample MST. Vuong (1989) showed how his method could be viewed as a natural generalization of the classical GLRT methodology. Golden (1998) noted that a simple and straightforward modification of Vuong's (1989) method called the DRMST (Discrepancy Risk Model Selection Test) is useful for constructing MST procedures for a wide variety of smooth goodness-of-fit functions.

Organization of this article. This article is organized in the following manner. First, some fundamental concepts associated with MSC and MST procedures are introduced and discussed. Second, the DRMST procedure is introduced and discussed. And third, simulation studies are provided to emphasize key similarities and differences between various MSC and MST procedures.

I. Definitions and Concepts for MSC and MST

Data Generating Processes and Probability Models

The data generating process. The *observed data* will be represented by a set of n vectors corresponding to n data points. The notation $\mathbf{X}_n = [\mathbf{x}(1), \ldots, \mathbf{x}(n)]$ will be used to denote a data sample of size n. The observed data is assumed to be generated by sampling from a population with a particular probability distribution. This specific probability distribution is called the *environmental distribution*. The mechanism for generating the independent and identically distributed (i.i.d.) data from the environmental distribution is called the *data generating process*.[1]

Probability model. A set of probability distributions is called a *probability model*.[2] Let p_θ be a probability mass (or density) function whose identity is specified by choosing a particular *parameter vector* θ. For example, if p_θ is a probability mass function, then $p_\theta(\mathbf{x}(i))$ is the probability mass assigned to observation $\mathbf{x}(i)$. A *parameter space* W for a probability model M is a set defined such that: p_θ is in M if and only if θ is in W. It will be implicitly assumed throughout this paper that p_θ is a sufficiently smooth function of θ.

Correctly specified, misspecified, and nested models. If the environmental distribution p^* is a member of a given probability model M, then M is a *correctly specified* model with respect to p^*. If the environmental distribution p^* is not a member of a given probability model M, then M is a *misspecified* model with respect to p^*. Let probability models F and G be subsets of the probability model M. If G is a subset of F, then the *reduced model* G is said to be *fully nested* in the *full model* F. Alternatively, suppose that $F \cap G = \emptyset$, then F and G are said to be *strictly non-nested*.

[1] The term "operating model" (Linhart & Zucchini, 1986; Zucchini, this issue) is an alternative term sometimes used to refer to the environmental distribution. This paper will always the term "model" to refer to a set of probability distributions.

[2] The term "approximating family of probability distributions" (Linhart & Zucchini, 1986; Zucchini, this issue) is an alternative name for the concept "probability model" which is used in this paper and the econometric literature.

Discrepancy Function Concepts

A discrepancy function measures the similarity between two probability distributions (Linhart and Zucchini, 1986; Zucchini, this issue). Let M be a probability model. Let Δ be a function that maps two probability distributions in M into a real number. Let p^* be the environmental distribution which is an element of M. Define another probability model F, which is a subset of M, which corresponds to the set of "proposed approximations" to the environmental distribution. A probability distribution p_{θ^*} that minimizes the quantity $\Delta(p^*, p_{\theta^*})$ subject to the constraint that $p_{\theta^*} \in F$ is called the *best approximating distribution* to p^* for F. The quantity $\Delta(p^*, p_{\theta^*})$ is called the *true model discrepancy* or *discrepancy due to approximation* (Linhart & Zucchini, 1986) for F with respect to p^*. The parameter vector θ^* is called an *optimal parameter vector*. It is implicitly assumed throughout this paper that the probability model F is chosen so that the optimal parameter vector for a given probability model F is always unique. Furthermore, it is assumed that Δ is a sufficiently smooth function of its arguments.

Model Selection Criterion Concepts

In general, the true model discrepancy $\Delta(p^*, p_{\theta^*})$ is not observable and must be estimated. One approach to estimating the true model discrepancy involves first estimating the environmental distribution p^* which is not known by the "empirical distribution" $\hat{p}_n^*$. The *empirical distribution* is a probability mass function (*directly derived* from a sample of size n) which assigns a probability to an observation that is equal to the relative frequency of that observation in the sample.

Definition of a model selection criterion. Now, using the empirical distribution function $\hat{p}_n^*$, choose a *minimum discrepancy parameter estimate* $\hat{\theta}_n$ corresponding to a probability distribution $p_{\hat{\theta}_n}$ in F that minimizes the *estimated model discrepancy* $\hat{\Delta}_n = \Delta(\hat{p}_n^*, p_{\hat{\theta}_n})$. The distribution $p_{\hat{\theta}_n}$ is called (in this paper) the *estimated best approximating distribution* for F with respect to $\mathbf{X}_n$. A model selection criterion $\mathcal{C}_n$ is an estimator of the true model discrepancy. The general

form for a model selection criterion is $\mathcal{C}_n = \hat{\Delta}_n + k_n$ where k_n is a "penalty term". Examples of typical model selection criteria will be discussed shortly.

Uniqueness assumption. It will be implicitly assumed throughout this paper that the observed data $\mathbf{X}_n$ and probability model F have the property that the estimated parameter vector $\hat{\theta}_n$ for F given $\mathbf{X}_n$ is always unique. This is a common assumption in statistical inference and simply means that there is sufficient information content in the data to select a particular distribution in F. For example, if F is a linear regression model with two free parameters and $\mathbf{X}_n$ consisted of only one data point, then the estimated best approximating distribution would not be unique because the linear regression model's free parameters are not uniquely determinable from the observed data. Bamber and van Santen (this issue) provide a detailed discussion of these issues.

Examples of Model Selection Criterion Functions

Gauss model selection criterion. Using the Gauss discrepancy function (Linhart and Zucchini, 1986; Zucchini, this issue), a Gauss model selection criterion is defined by the formula:

$$\mathcal{C}_n = \sum_{i=1}^{n} (\hat{p}_n^*(\mathbf{x}(i)) - p_{\hat{\theta}_n}(\mathbf{x}(i)))^2.$$

Kolmogorov model selection criterion. An alternative to the Gauss discrepancy function is the Kolmogorov discrepancy function (see Linhart and Zucchini, 1986, p. 18) which gives rise to a Kolmogorov model selection criterion:

$$\mathcal{C}_n = max\{|\hat{p}_n^*(\mathbf{x}(1)) - p_{\hat{\theta}_n}(\mathbf{x}(1))|, \ldots, |\hat{p}_n^*(\mathbf{x}(n)) - p_{\hat{\theta}_n}(\mathbf{x}(n))|\}$$

where the quantity $max\ S$ denotes the largest element in S. Almost all of the theoretical results in this paper are *not applicable* to the Kolmogorov discrepancy function since this discrepancy function is not a smooth function of its arguments.

Log-likelihood model selection criterion. The log-likelihood discrepancy function (e.g., White, 1982; Golden, 1996) results in a model selection criterion

corresponding to the probability model whose estimated best approximating distribution maximizes the likelihood of the observed data. The log-likelihood model selection criterion is defined as:

$$\mathcal{C}_n = -(1/n)\sum_{i=1}^{n} log(p_{\hat{\theta}_n}(\mathbf{x}(i)).$$

Note that the minimum discrepancy parameter estimate $\hat{\theta}_n$ in this case is called a *maximum likelihood estimate* when the model is correctly specified.

Penalized log-likelihood model selection criterion. A number of researchers (e.g., Akaike, 1973; Schwarz, 1978; Sin and White, 1996) have proposed various modifications to the log-likelihood discrepancy function that effectively penalize models containing excessive free parameters. Such penalties are typically introduced using the model selection criterion:

$$\mathcal{C}_n = -(1/n)\sum_{i=1}^{n} log(p_{\hat{\theta}_n}(\mathbf{x}(i)) + k_n$$

where $\hat{\theta}_n$ is a q-dimensional minimum discrepancy parameter vector estimate. The number k_n is called k_n is an optional *penalty term*. For example, if $k_n = q/n$, then k_n is called the AIC (Akaike Information Criterion) penalty term. If $k_n = (q/2)(log(n)/n)$, then the SIC (Schwarz Information Criterion) penalty term is obtained. The BIC (Bayes Information Criterion) penalty term, which is also widely used, is effectively equivalent to the SIC penalty term (Kass and Waterman, 1995). Note that the parameter estimate $\hat{\theta}_n$ is the minimum discrepancy estimator as usual.

As the sample size n becomes large, the large sample distribution of the random variable estimated by $\mathcal{C}_n$ becomes less dependent upon the penalty term k_n. On the other hand, for a fixed sample size, *important advantages* of the "bias" introduced by the penalty term k_n are obtained. The other papers in this special issue (e.g., Bozdogan, Forster, Myung, Zucchini) discuss some of those advantages (also see Akaike, 1973; Bozdogan, 1987; Kass and Wasserman, 1995; Schwarz, 1978; Sin and White, 1996 for relevant discussions).

Discrepancy risk model selection criterion. The above model selection criteria can be represented in a more general unified framework by introducing

the concept of a discrepancy risk model selection criterion. The discrepancy risk model selection criterion results in a model selection criterion where the probability model is chosen whose best approximating distribution minimizes expected loss.

A *discrepancy loss function* c_θ is a function that maps a particular probability distribution, p_θ, and an observation, $\mathbf{x}(i)$, from the data sample into a real number. The quantity $c_\theta(\mathbf{x}(i))$ is defined as the *true discrepancy loss* for choosing the probability distribution p_θ as the best approximating distribution in the probability model given the data point $\mathbf{x}(i)$ is observed.

In practice, the true discrepancy loss is not directly observable because the best approximating distribution, p_θ, in the probability model is not known. Thus, the true discrepancy loss must be estimated. Accordingly, the quantity $c_{\hat{\theta}_n}(\mathbf{x}(i))$ is defined as the *estimated discrepancy loss* for choosing the estimated best approximating distribution in the probability model given the data point $\mathbf{x}(i)$ is observed.

Given the above definition of a discrepancy loss function, the discrepancy risk model selection criterion is defined by the formula:

$$\mathcal{C}_n = \hat{\Delta}_n + k_n$$

where the estimated model discrepancy

$$\hat{\Delta}_n = (1/n) \sum_{i=1}^{n} c_{\hat{\theta}_n}(\mathbf{x}(i))$$

and where k_n is an optional penalty term. The penalty term k_n is a realization of the random variable $\tilde{k}_n$ which is assumed to have the property that $\sqrt{n}\tilde{k}_n$ converges to zero in probability as $n \to \infty$. The parameter estimate $\hat{\theta}_n$ is the minimum discrepancy estimator as usual.

The MSC Procedure

Given a set of probability models and an environmental distribution p^*, one would ideally like to select the probability model with the smallest true model discrepancy with respect to p^*. In practice, however, p^* is not observable and the researcher usually only has the observed data $\mathbf{X}_n$. Thus, the researcher uses a *Model Selection*

Criterion (MSC) to select the probability model which has the smallest estimated model discrepancy with respect to $\mathbf{X}_n$. The estimated model discrepancy will usually be a good approximation to the true model discrepancy if the sample size is sufficiently large (see Appendix A of Linhart and Zucchini, 1986, for specific details).

Using the notation of the previous sections, let $\hat{\Delta}_F$ be the estimated model discrepancy for some probability model F. Let $\hat{\Delta}_G$ be the estimated model discrepancy for some probability model G. Let $\mathcal{C}_n^F = \hat{\Delta}_F + k_n^F$ and $\mathcal{C}_n^G = \hat{\Delta}_G + k_n^G$. Let

$$\hat{\delta}_n = \mathcal{C}_n^F - \mathcal{C}_n^G = (\hat{\Delta}_F - \hat{\Delta}_G) + (k_n^F - k_n^G)$$

be defined as the *estimated average between-model discrepancy error*. The MSC procedure chooses probability model F if the estimated average between-model discrepancy error $\hat{\delta}_n < 0$ and chooses probability model G if $\hat{\delta}_n > 0$. If (in the extremely rare event) that $\hat{\delta}_n = 0$, then neither probability model is selected.

Note the quantity $\hat{\delta}_n$ is actually the observed value of a random variable $\tilde{\delta}_n$. The random variable $\tilde{\delta}_n$ will take on a different value *for a fixed sample size* n depending upon the *particular sample* of size n generated from the environmental distribution.

Using the Strong Law of Large Numbers, it follows that $\tilde{\delta}_n$ converges almost surely to the *true average between-model discrepancy error*. Thus, for large sample sizes, the MSC procedure (which is based upon the *estimated model discrepancy*) is approximately equivalent to choosing the probability model whose model has the smallest *true model discrepancy* as the sample size n becomes large.

II. A Large Sample Pair-wise Model Selection Test

As the sample size becomes large, the MSC procedure selects the probability model with a smallest true model discrepancy. For a given *fixed* sample size, the *estimated* average between-model discrepancy error is a random variable with a particular probability distribution. In this section, a large sample MST is introduced for

testing the *MST null hypothesis* that the *true* average between-model discrepancy error is significantly different from zero at some chosen significance level. If the MST null hypothesis is rejected, then one concludes that at the chosen significance level there is sufficient evidence to conclude that the probability model with the smallest *estimated* model discrepancy (i.e., the MSC procedure rule) also has the smallest *true* model discrepancy for a *fixed* sample size. If the MST null hypothesis is not rejected, then one concludes that at the chosen significance level there is not sufficient evidence to conclude one probability model has a smaller *true* model discrepancy than the other.

This section is organized into two subsections. In the first subsection, a relatively simple large sample MST for comparing two (possibly misspecified) probability models which are strictly non-nested is described based upon the methods of Linhart (1988) and Vuong (1989). This MST for strictly non-nested probability models is then noted to be inappropriate for situations where the probability models are not strictly non-nested. In the second subsection, a relatively complex large sample MST for comparing possibly misspecified and non-nested probability models under a wide variety of conditions is described based upon the method of Vuong (1989; see Golden, 1998, for further details).

The Strictly Non-nested Models Case

A MST for strictly non-nested models. Linhart (1988; also see Shimodaira, 1997) and Vuong (1989) have proposed the following *strictly non-nested* MST for testing the null hypothesis that two strictly non-nested probability models have the same true model discrepancy. Let $c_{\theta_F}(\mathbf{x}(i))$ be the true discrepancy loss for probability model F given observation $\mathbf{x}(i)$. Similarly, let $c_{\theta_G}(\mathbf{x}(i))$ be the true discrepancy loss for selecting probability model G given observation $\mathbf{x}(i)$. Let the quantity

$$\tilde{\epsilon}_i = c_{\theta_F}(\tilde{\mathbf{x}}(i)) - c_{\theta_G}(\tilde{\mathbf{x}}(i))$$

be called the *true between-model discrepancy observation error*.

In practice, c_{θ_F} must be estimated using the estimated best approximating distribution for model F since the environmental distribution which generated the observed data is not directly observable. Accordingly, the estimated value of $c_{\theta_F}(\mathbf{x}(i))$, $c_{\hat{\theta}_{F,n}}(\mathbf{x}(i))$, is used to compute the *estimated between-model discrepancy observation error*, $\hat{e}_i$. An explicit formula for the estimated between-model discrepancy error is given by:

$$\hat{e}_i = c_{\hat{\theta}_{F,n}}(\mathbf{x}(i)) - c_{\hat{\theta}_{G,n}}(\mathbf{x}(i)). \tag{1}$$

Note that

$$\tilde{\delta}_n = (1/n)\sum_{i=1}^{n} \tilde{e}_i + (\tilde{k}_n^F - \tilde{k}_n^G)$$

is approximately the average of all n between-model discrepancy observation errors (since $\sqrt{n}(\tilde{k}_n^F - \tilde{k}_n^G) \to 0$ in probability as $n \to \infty$). The Central Limit Theorem states that the square root of n multiplied by the average of a set of n i.i.d. zero-mean random variables and common positive variance σ^2 will converge in distribution to a Gaussian random variable with mean zero and variance σ^2. Assume that the variance of $\tilde{\epsilon}_i$ which will be denoted as σ_ϵ^2 is strictly positive. The assumption $\sigma_\epsilon^2 > 0$ is referred to as the *variance assumption*. Vuong (1989, Lemma 4.1) showed for the log-likelihood case that an equivalent way of stating the variance assumption is that $F \cap G = \emptyset$ (i.e., the probability models F and G are strictly non-nested).

Given the variance assumption and noting that the random variables $\tilde{e}_1, \ldots, \tilde{e}_n$ are asymptotically i.i.d. (because the observations are i.i.d. and $\tilde{e}_i \to \tilde{\epsilon}_i$ almost surely), it follows[3] from the Central Limit Theorem that $\sqrt{n}\tilde{\delta}_n$ converges in distribution to a Gaussian distribution with mean zero and variance σ_ϵ^2 if the MST null hypothesis is true. Note that if the MST null hypothesis is false that $|\sqrt{n}\tilde{\delta}_n| \to \infty$ almost surely as the sample size $n \to \infty$.

[3]In this overview, a number of technical details and assumptions have been omitted for expository reasons. For a mathematically rigorous version of these arguments, please see Golden (1998) for the general case discussed here or Vuong (1989) for the log-likelihood discrepancy function case.

The above observations can be used to construct a slightly more general MST for some chosen significance level α provided that one assumes that the variance assumption is true. First, estimate the (presumably strictly positive) variance σ_ϵ^2 by the quantity:

$$\hat{\sigma}_\epsilon^2 = (1/n)\sum_{i=1}^{n} \hat{e}_i^2 - (E[\tilde{\epsilon}_i])^2 \tag{2}$$

where E denotes the expectation operator. Equation (2) then reduces to:

$$\hat{\sigma}_\epsilon^2 = (1/n)\sum_{i=1}^{n} \hat{e}_i^2 \tag{3}$$

given the null hypothesis $E[\tilde{\epsilon}_i] = 0$ (i.e., that the true model discrepancies are equal). Second, compute

$$Z_{obs} = \frac{\sqrt{n}\hat{\delta}_n}{\hat{\sigma}_\epsilon}. \tag{4}$$

Third, use either a table of Z-scores or a computer software program to compute Z_α which is the probability α that the magnitude of a normally distributed random variable exceeds the value of Z_α. Fourth, if $|Z_{obs}| > Z_\alpha$, reject the MST null hypothesis at the α significance level. Otherwise, if $|Z_{obs}| \leq Z_\alpha$, do not reject the MST null hypothesis at the α significance level.

Example non-nested MST problem. To illustrate these concepts, consider the following non-nested model selection test example. Suppose that the observed data set of sample size $n = 4$ is defined as $\{2, 3, 10, 3\}$ (i.e., $\mathbf{x}(1) = 2$, $\mathbf{x}(2) = 3$, $\mathbf{x}(3) = 10$, and $\mathbf{x}(4) = 3$). Let F be the set of all univariate Gaussian density functions whose variance is equal to 1. Let G be the set of all univariate Gaussian density functions whose variance is equal to 2. The problem which is to be solved is to construct a MST using a log-likelihood model selection criteria for deciding whether or not to reject the null hypothesis $H_0 : \Delta_F = \Delta_G$. Or, in other words, the problem involves testing the null hypothesis the best approximating distribution from the probability model F or the best approximating distribution from the probability model G are equally "distant" from the environmental distribution.

Solution to example non-nested MST problem. Since $F \cap G = \emptyset$, this means that the models are strictly non-nested so the model selection test for strictly non-nested models is appropriate.

Since F is the set of univariate Gaussian density functions and the log-likelihood discrepancy function will be used, then the formula for $c_{\hat{\theta}_{F,n}}(\mathbf{x}(i))$ is given by the expression:

$$c_{\hat{\theta}_{F,n}}(\mathbf{x}(i)) = -log(p_F(\mathbf{x}(i)) \tag{5}$$

where

$$p_F(\mathbf{x}(i)) = \frac{exp(-(\mathbf{x}(i) - \hat{m}_n)^2/2)}{\sqrt{2\pi}},$$

is the univariate Gaussian probability density function with a known variance equal to 1. The large sample maximum likelihood estimate $\hat{m}_n$ is simply the average of the observations and is given by the formula $\hat{m}_n = (2 + 3 + 10 + 3)/4 = 4.5$.

Similarly, the formula for $c_{\hat{\theta}_{G,n}}(\mathbf{x}(i))$ is given by the expression:

$$c_{\hat{\theta}_{G,n}}(\mathbf{x}(i)) = -log(p_G(\mathbf{x}(i)) \tag{6}$$

where

$$p_G(\mathbf{x}(i)) = \frac{exp(-(\mathbf{x}(i) - \hat{m}_n)^2/4)}{2\sqrt{\pi}},$$

is the univariate Gaussian probability density function with a known variance equal to 2. The large sample maximum likelihood estimate $\hat{m}_n$ is simply the average of the observations and is given by the formula $\hat{m}_n = (2 + 3 + 10 + 3)/4 = 4.5$.

Now, evaluate Equation (5) at each of the four data points to obtain the estimated discrepancy loss for each data point under model F. For the data points $\{2, 3, 10, 3\}$, these estimated discrepancy losses are $\{4.04, 2.04, 16.04, 2.04\}$ respectively. Similarly, using Equation (6), the estimated discrepancy losses for $\{2, 3, 10, 3\}$ are $\{2.83, 1.83, 8.83, 1.83\}$ respectively. Thus, using Equation (1), the estimated between-model discrepancy errors for $\{2, 3, 10, 3\}$ are $\{1.22, 0.22, 7.22, 0.22\}$ respectively.

The mean and the standard deviation of the estimated between-model discrepancy errors are $m = 2.21$ and $s = 2.53$ respectively. Thus, using Equation (4), $Z_{obs} = (2.21/2.53)\sqrt{4} = 1.75$. Using $\alpha = 0.05$, from a table of Z-scores, it follows that $Z_\alpha = 1.96$. Since $Z_{obs} = 1.75$, it follows that $|Z_{obs}|$ is less than Z_α so the null hypothesis that $\Delta_F = \Delta_G$ is not rejected. The researcher would conclude that: (i) either the two probability models provide equally effective fits to the environmental distribution using a log-likelihood model selection criteria, or (ii) the non-nested MST has insufficient statistical power to reject the null hypothesis for the given data sample. Note that $\mathcal{C}_n^F = 6.04$ and $\mathcal{C}_n^G = 3.83$ but according to the non-nested MST at the $\alpha = 0.05$ significance level, there is not sufficient evidence to reject the null hypothesis $H_0 : \Delta_F = \Delta_G$. This differs from a MSC conclusion based upon the same log-likelihood model selection criteria which would select model G instead of model F since $\mathcal{C}_n^G < \mathcal{C}_n^F$.

Could the variance assumption ever be incorrect? The variance assumption presented in the previous section can not be taken lightly since in many important and typical situations where the MST might be applied, the variance assumption is *incorrect*! Both Vuong (1989) and Shimodaira (1997) have emphasized very clearly that, in general, *the variance assumption is incorrect when the two probability models are not strictly non-nested.* In particular, Vuong (1989) explicitly showed this by characterizing the distribution of $\tilde{\delta}_n$ for log-likelihood discrepancy functions under very general conditions.

For example, using Wilk's (1938) GLRT, it follows that (when two probability models are nested, the full model is correctly specified, and a log-likelihood discrepancy risk function is used) the distribution of $\sqrt{n}\tilde{\delta}_n$ is *not* Gaussian. Rather, $2n\tilde{\delta}_n$ has a Chi-square distribution with degrees of freedom equal to the difference in free parameters between the two probability models (see Vuong, 1989, or Golden, 1995, 1996, for relevant reviews).

What does the variance assumption really mean? The variance of σ_ϵ^2 given by the formula:

$$\sigma_\epsilon^2 = E[\tilde{\epsilon}_i^2] - (E[\tilde{\epsilon}_i])^2$$

where E denotes expectation taken with respect to the environmental distribution which generated the observed data. Under the null hypothesis that the true model discrepancies are equal and using the definition of ϵ_i we have:

$$\sigma_\epsilon^2 = E[(c_{\theta_F}(\tilde{\mathbf{x}}(i)) - c_{\theta_G}(\tilde{\mathbf{x}}(i))^2].$$

Inspection of this expression for σ_ϵ^2 reveals that $\sigma_\epsilon^2 = 0$ if and only if c_{θ_F} and c_{θ_G} are almost surely identical (see Lemma 4.1 of Vuong, 1989, for a similar argument for the specialized log-likelihood discrepancy function case).

In other words, the *true* between-model discrepancy observation error is equal to zero for all possible observations if and only if the variance assumption ($\sigma_\epsilon^2 = 0$) is false. It should be emphasized that the researcher will only observe the *estimated* between-model discrepancy observation error for every possible observation and these estimated between-model discrepancy errors will typically *not* be exactly equal to zero. Thus, even if $\hat{\sigma}_\epsilon^2$ is strictly positive, the quantity σ_ϵ^2 (which is *estimated* by $\hat{\sigma}_\epsilon^2$) may be exactly equal to zero.

To illustrate these ideas, consider the typical situation where a probability model F is fully nested within a probability model G. Also assume that the data is generated from some environmental probability distribution p^* which is an element of model F. Also assume that log-likelihood discrepancy functions are used. Since the best approximating distribution is shared by both probability models, the true between-model discrepancy observation error will be exactly equal to zero for every observation (i.e., the variance assumption is false).

The DRMST (Discrepancy Risk Model Selection Test)

Overview of the DRMST. The Discrepancy Risk Model Selection Test (DRMST) is a two stage MST designed to extend the simple large sample MST for strictly non-nested model comparison to the more general case where the two probability models are possibly misspecified or non-nested.

In the first stage of the DRMST, one uses a statistical test called the *variance test* in order to decide whether or not to reject the null hypothesis that the variance assumption (i.e., $\sigma_\epsilon^2 = 0$) is false. If one rejects the null hypothesis that the variance

assumption is false (i.e., the assumption that σ_ϵ^2 is strictly positive holds), then the second stage MST for strictly non-nested models can be used to decide whether or not to reject the null hypothesis that the true average between-model discrepancy error is equal to zero. It can be shown (see Vuong, 1989, p. 321 for a discussion of the log-likelihood case; the more general case is based upon similar arguments) that if the significance levels of the two component statistical tests of the DRMST (i.e., the variance test and the MST for strictly non-nested models) are both equal to α, the resulting two-stage statistical test will asymptotically have a significance level less than or equal to α.

Figure 1 illustrates the basic concept and logic of the DRMST.

Insert Figure 1 About Here

Golden (1998) introduces and describes the DRMST in greater detail using the method of Vuong (1989). Golden's (1998) analysis is best viewed as an almost immediate extension of the method of Vuong (1989) which is applicable to comparing possibly misspecified and non-nested models using a log-likelihood risk discrepancy function. Vuong's (1989) research, in turn, was largely inspired by work in the area of hypothesis testing in the presence of model misspecification (White, 1982; White, 1994; see Golden, 1995, for a review).

How does one implement the DRMST? Although the large sample MST for strictly non-nested models is easily implemented using most standard statistical software packages, the variance test is relatively complicated. Implementation of the variance test requires explicit computation of the matrix first derivatives and matrix second derivatives of the discrepancy loss function with respect to the probability model parameters for both probability models in order to compute the "weights" for a special random variable called a "weighted chi-square random variable". Critical values of the weighted chi-square random variable must then be computed using specialized computer software. Golden (1995) reviews in detail Vuong's (1989) implementation of the DRMST for log-likelihood discrepancy

risk functions, while Golden (1998) provides explicit formulas for a much more general class of discrepancy functions. For the special case of linear, logistic, and multinomial logit regression modeling with weighted and unweighted log-likelihood discrepancy functions, easy-to-use documented commercial software tools are now becoming available (e.g., see the *CCR Modeling System* software under development by Martingale Research Corporation, 1998).

Properties of the DRMST. The DRMST has a number of key properties. First, it is applicable to a large class of "sufficiently smooth" discrepancy functions which includes most (but not all) popular probability models and discrepancy functions. Second, the DRMST can only be used to compare two probability models at a time. Third, the DRMST can be used as a tool for using a given MSC procedure to decide which of two probability models "best-fits" the underlying data generating process, and transforming that MSC procedure into a large sample statistical test for deciding if the observed differences in "goodness-of-fit" are statistically significant. Fourth, the DRMST is a natural generalization of the classical Wilk's (1938) GLRT (see Vuong, 1989, for specific details). Fifth, unlike the GLRT, the models may be non-nested, partially nested, or (as in the GLRT) fully nested. Sixth, unlike the GLRT, it is not required that at least one of the probability models is correctly specified with respect to the underlying data generating process. And seventh, like the GLRT, the DRMST is a large sample statistical test whose behavior for small samples has not been investigated and is likely to vary considerably in different applications for a given sample size. This latter point implies that the large sample approximations should be checked when the DRMST is applied to new situations against an alternative "simulation method" such as the large sample non-parametric boot strap methodology (see Golden, 1995, and Zucchini, this issue).

III. Simulation Studies: Comparisons of MSC and MST Methods

In order to illustrate the ideas presented in the previous sections, this section discusses some simulation studies. In Simulation Study 1, the MSC and MST

procedures were used to decide if a particular probability model fits the data better than the model within which it is fully nested. Simulation Study 2 compared the MSC procedure to the MST procedure in the case where the two probability models are non-nested.

Common Features of the Simulation Studies

Logistic regression models and discrepancy loss functions. The outcome random variable $\tilde{x}$ for a logistic regression model is assumed to take on only two possible values: zero or one. The notation $p(x)$ will be used to denote the probability the outcome variable is equal to one given an observed value x of the predictor variable. In a simple logistic regression model,

$$p(x) = \mathcal{S}(mx + b)$$

where $\mathcal{S}(x) = 1/(1 + exp(-x))$, m is the *slope* parameter and b is the *intercept* parameter. All simulation studies were based upon the log-likelihood discrepancy loss function which is defined for a simple logistic regression model by the formula:

$$c(x) = -[xlog(p(x)) + (1 - x)log(1 - p(x))].$$

Data generating process. The data in all of these simulation studies was generated from a particular environmental probability distribution which will be referred to as p^*. The predictor variable value was a random number uniformly distributed on the interval [-0.5,+0.5]. The binary outcome variable was constrained to take on the value of 0 with probability 0.4 and 1 with probability 0.6 (i.e., $p^*(\tilde{x} = 1) = 0.6$ and $p^*(\tilde{x} = 0) = 0.4$).

Notice that the outcome variable's value was generated *independently* of the value of the "predictor" variable. In most modeling problems, a parameter of the probability model may have an estimated value that is not significantly different from zero. Such an "irrelevant" model parameter is usually eliminated from the probability model because there is no reason to reject the null hypothesis that its optimal value is equal to zero. Accordingly, the environmental distribution p^* in this simulation study is designed to model the situation where the null hypothesis

an optimal model parameter is zero is in fact true. In such a situation, the large sample probability distribution of the parameter estimate about the value of zero can be estimated for a given sample size. More specifically, in the simulation studies reported here, it is assumed that p^* is defined by a simple logistic regression model whose slope parameter m is *exactly equal to zero.*

Data samples. Data samples of three different sizes were generated from p^* in order to explore how the performance of the MSC and MST procedures discussed here varied as the sample size n becomes large. This is an important issue since the MSC and MST procedures described here are all large sample methods. The sample sizes were $n = 10$, $n = 100$, and $n = 1000$ so that a *sample* of size $n = 10$ consisted of 10 data records such that each data record consisted of the value of a predictor variable value and an outcome variable value. For each of the three sample sizes, 100 samples of size n were generated.

Simulation Study 1: Nested Models

In this simulation study, two nested probability models were compared with respect to data samples generated from an environmental distribution which is an element of the full probability model. The data was generated from an environmental distribution common to both probability models. Thus, the reduced model provides the most "parsimonious" explanation of the data (see Myung, this issue). On the other hand, if one is interested in the related (but distinct) problem of identifying situations where both models appear to provide approximately equally well fits to the underlying data generating process, then it would be desirable to have a methodology such as the MST methodology specifically designed to test the null hypothesis both models are "equally effective".

Methodology. In the first set of simulation studies, parameters for two nested logistic regression models were estimated using maximum likelihood estimation. In particular, the full logistic regression model F was defined by:

$$F = \{p : p(x) = \mathcal{S}(mx + b), m \in \mathcal{R}, b \in \mathcal{R}\},$$

while the reduced logistic regression model G was defined by:

$$G = \{p : p(x) = \mathcal{S}(mx + b), m = 0, b \in \mathcal{R}\}.$$

Or, in other words, the full model F consisted of probability distributions of the form $p(x) = \mathcal{S}(mx + b)$, while the reduced model G consisted of probability distributions of the form $p(x) = \mathcal{S}(b)$. Also note that the true model discrepancy for G must always be less than or equal to the true model discrepancy for F since every distribution in G can be represented by some distribution in F.

A log-likelihood MSC procedure was used to select either the full or reduced probability model for the three sample sizes of $n = 10$, $n = 100$, and $n = 1000$. The number of times the full probability model was chosen by the MSC was recorded. Similarly, the number of times the reduced probability model was chosen by the MSC was recorded. In addition to a log-likelihood MSC procedure, two other MSC procedures (log-likelihood MSC with an AIC penalty term and log-likelihood MSC with a BIC/SIC penalty term) and the GLRT MST procedure (with a α-significance level) were used.

Note that the GLRT MST procedure is formally equivalent to the variance test (see Vuong, 1989, Corollary, 7.5; see Golden, 1995, for a review). However, since the models were fully nested and the full model is correctly specified, the calculations were considerably simplified and Wilk's (1938) original GLRT chi-square statistic could be used. In particular, let $\chi^2_\alpha(1)$ be defined such that the probability that a chi-square random variable with 1 degree of freedom obtains a value greater than $\chi^2_\alpha(1)$ is equal to α. Then, at the α-significance level, reject the null hypothesis both models are equally effective if $2n|\hat{\delta}_n| > \chi^2_\alpha(1)$. When the null hypothesis is in fact rejected, one concludes that the full model provides a significantly better fit to the data than the reduced model.

Results and discussion. Table 1 shows the simulation results for the three MSC procedures and the GLRT MST procedure. For example, Table 1 shows that out of 100 computer generated samples of size $n = 1000$, the MSC procedure selects the least parsimonious model 78 out of 100 samples (and never selects the most parsimonious model). The remaining 22 samples generated "ties" in the

sense that the estimated average between-model discrepancy error $\hat{\delta}_n$ was exactly equal to zero (because parameter estimates and model discrepancy estimates in all simulations were rounded to six digit precision).

Notice that the MSC procedure without a penalty term tends to select the less parsimonious full model because of an overfitting phenomenon. This problem is partially corrected through the use of the MSC procedures involving the penalty terms with the BIC/SIC penalty term showing an advantage over the AIC penalty term in this particular set of simulation runs. The non-nested MST has a strong tendency to select neither probability model and thus indicates both probability models fit the underlying data generating process equally effectively.

Simulation Study 2: Non-nested Models

The previous simulation study considered the fully nested and correctly specified case. Simulation Study 2 considered a situation where the two probability models are non-nested and one of the probability models is misspecified. Thus, there is a "correct" and "wrong" model selection decision in Simulation Study 2 (unlike Simulation Study 1 where both probability models contained the environmental distribution).

Methodology. In simulation study 2, one logistic regression model F was assumed to have a slope parameter m but no intercept parameter b, so that a distribution in F was assumed to have the form: $p(x) = \mathcal{S}(mx)$. The other logistic regression model G was assumed to have an intercept parameter b but no slope parameter m, $p(x) = \mathcal{S}(b)$. The two logistic regression models are therefore non-nested.

Note both models have exactly the same number of free parameters (each model has exactly 1 free parameter), thus standard methods such as the AIC or BIC model selection criteria have no effect in this situation upon the model selection process. However, other types of penalty terms based upon the functional form of the approximating probability model can have effects in this type of situation. For example, Zucchini (this issue; also see Linhart and Zucchini, 1986, Appendix

A1) discusses a generalization of the AIC model selection criteria appropriate in the presence of model misspecification which will have an effect on the MSC model selection process even when both models have the same number of free parameters.

Results and discussion. Table 2 compares the results of a log-likelihood MSC procedure with a log-likelihood MST procedure for this non-nested model selection problem. The MSC procedure tends to select the wrong probability model because both models have the same numbers of free parameters and the model with the slope parameter (i.e., the probability model which does not contain the environmental distribution) has a tendency to fit the noise in the data set in this special situation. Note that when the slope parameter is set equal to the value of zero, the resulting probability distribution $p(\tilde{x} = 1) = 0.5$ is very close but still distinct from the environmental distribution $p(\tilde{x} = 1) = 0.6$. The MST procedure, on the other hand, behaves very conservatively and refuses to reject the null hypothesis that both models fit the underlying data generating process $p(\tilde{x} = 1) = 0.6$ equally well.

Summary

This paper has discussed a particular large sample MST for testing the null hypothesis two models have the same true model discrepancy. Or, less formally, the MST may be used to decide if an observed difference in "estimated goodness-of-fit" between two probability models is significantly different from zero. The MST described here is a natural extension of the well known GLRT but is also applicable in situations where one or both of the models may be misspecified and the models may or may not be nested.

Illustrative simulation studies of both the MSC and MST in the presence and absence of penalty terms (such as the AIC and BIC/SIC terms) emphasized that the MST results in a relatively conservative decision rule where the option of deciding that one model fits the underlying distribution better than the other model is not invoked until a threshold (determined by the significance level of the MST) has been reached.

Finally, explicit formulas for using the MST described here for log-likelihood discrepancy functions may be found in Vuong (1989; also see Golden, 1995). Golden (1998) provides explicit formulas for a wide class of smooth discrepancy functions by exploiting the methods of Vuong (1989). Commercial computer software for implementing the large sample MST for linear, logistic, and multinomial logit regression models is also available (Martingale Research, 1998). In conclusion, the MST approach described here is an accessible and useful large sample statistical test that can be applied in a great variety of important situations.

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.

Bamber, D. & J. P. H. van Santen (this issue). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology.*

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

Bozdogan, H. (this issue). Akaike information criterion and recent developments in informational complexity. *Journal of Mathematical Psychology.*

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, Series B,*, *24*, 323–353.

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

Forster, M. R. (this issue). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology.*

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). *A Large Sample Expected Loss Model Selection Test.* School of Human Development, University of Texas at Dallas, Richardson, TX: Unpublished manuscript.

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.

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: Wiley.

Martingale Research Inc. (1998). *CCR (Constrained Categorical Regression) Modeling Computer Software*. 1485 Richardson, Suite 110, Richardson, TX.

Myung, I. J. (this issue). The importance of complexity in model selection. *The Journal of Mathematical Psychology*.

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

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.

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

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

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

Zucchini, W. (this issue). An introduction to model selection. *The Journal of Mathematical Psychology*.

Author Notes

I am grateful to the School of Human Development at the University of Texas at Dallas for supporting this research. I would also like to thank Steven Henley, In Jae Myung, two anonymous reviewers, and especially Malcolm Forster for their valuable comments and feedback on an earlier version of this article.

Please address reprint requests and correspondence to: Richard M. Golden (golden@utdallas.edu), Cognition and Neuroscience Program, School of Human Development (GR4.1), University of Texas at Dallas, Box 830688, Richardson, Texas 75083-0688.

Table 1: Simulation Comparisons of MSC and MST (Nested Case)

Methodology	Model Selected	Sample Size		
		n = 10	n = 100	n = 1000
MSC	Full	97	90	78
	Reduced	0	0	0
MSC+AIC	Full	25	9	11
	Reduced	75	91	89
MSC+BIC/SIC	Full	14	2	1
	Reduced	86	98	99
MST	Full	7	4	4
	Reduced	0	0	0

Table 2: Simulation Comparisons of MSC and MST (Non-Nested Case)

Methodology	Model Selected	Sample Size		
		n = 10	n = 100	n = 1000
MSC	Wrong	97	90	78
	Correct	0	0	0
MST	Wrong	7	4	4
	Correct	0	0	0

Figure Captions

Figure 1. DRMST overview. If first stage null hypothesis *Variance assumption is false* is rejected using variance test, then test second stage null hypothesis *Both models have same true model discrepancy* using non-nested MST. If second stage null hypothesis is rejected, select model using MSC procedure. If either first or second stage null hypothesis is accepted, do not select either model.