Making Correct Statistical Inferences Using a Wrong Probability Model

RICHARD M. GOLDEN

University of Texas at Dallas

Large sample methods for estimating the variance of parameter estimates for hypothesis-testing purposes and statistical test for model selection when the statistical model is wrong (i.e., misspecified) are reviewed. A parallel distributed processing (PDP) statistical model for analyzing categorical time series data is then proposed, and a theorem establishing when the quasi-maximum likelihood estimates of the model are unique is stated and proved. Analysis of Golden *et al.*'s (1993, in the *Proceedings of the* 14th *Annual Conference of the Cognitive Science Society* (pp. 487–491). Hillsdale, NJ: Erlbaum) categorical time-series data with respect to the proposed PDP model showed that White's asymptotic statistical theory yielded results which were more consistent with boot-strap estimates than classical methods of statistical inference. © 1995 Academic Press, Inc.

Ideally, a statistical analysis should be as "model-independent" as possible, making a minimal number of assumptions about the nature of the data generating process. Such an analysis is exemplary of the classical data analysis approach. Unfortunately, however, such "ideal" analyses are usually not practical because (i) the data may be too "messy" to uniquely identify the nature of the data generating process, (ii) the experimenter may not be able to "adequately control" critical factors influencing the data generating process, and (iii) it is impractical for the experimenter to collect a sufficient amount of data.

In addition, consider the case of the experimenter who already has formulated a highly restricted class of alternative psychological theories regarding how the data were generated. The experimenter acknowledges that each theory in this class may not be correct in its entirety, but is simply interested in which theory is "closest" to the truth. Such models presumably reflect the experimenter's biases regarding the data and naturally incorporate those biases directly into the data analysis. Thus, the statistical model is a formal instantiation of the experimenter's psychological theories. Some examples of this strategy include the multinomial models discussed by Riefer and Batchelder (1988), linear causal path analysis models (e.g., Bentler, 1980), item response theory (e.g., Lord, 1980) for assessing latent subject ability parameters, categorical time-series analysis models for analyzing social interaction data (Allison & Liker, 1982), hidden Markov models for speech recognition (e.g., Levinson *et al.* 1988), or the learning model described by Falmagne *et al.* (1990).

The classical approach to hypothesis testing in complex statistical models is traditionally a two step procedure. First, the null hypothesis that the assumed statistical model could have generated the observed data is tested. If this null hypothesis is accepted, then the experimenter proceeds to the second step of the analysis. In the second step, specific planned comparisons involving relations among the estimated parameters are tested. Because the classical approach is based upon estimating the "true" parameters and their associated variances and covariances, tests of specific planned comparisons are permitted only if the model "fits" the data.

The two step classical approach to hypothesis testing is problematic because the first step is based upon accepting the null hypothesis that the assumed statistical model actually generated the observed data. If, for example, the data were simply random noise, then the null hypothesis would be accepted because of the large standard errors associated with the parameter estimates of the model. On the other hand, if the model was a truly superb model of the data generating process with negligible minor structural defects, the null hypothesis could be rejected if "too much data" is collected.

A probability model is a collection of probability distribution functions which are indexed by the probability model's parameters. If the data generating process is not contained in a particular probability model, then that probability model is referred to as a "misspecified model." To avoid the presence of some forms of model misspecification,

I am grateful to Bill Batchelder, A. A. J. Marley, Stephen Hayes, and especially two anonymous reviewers for their comments on earlier versions of this article. Some of the ideas regarding a connectionist interpretation of a probabilistic category choice model evolved from discussions with Dr. David Rumelhart while the author was a National Institute of Health postdoctoral fellow at Stanford University in 1989–1990.

Correspondence and requests for reprints should be sent to Richard Golden, University of Texas at Dallas, School of Human Development, GR41, Box 830688, Richardson, Texas, 75083-0688. E-mail address: golden@utdallas.edu.

0022-2496/95 $6.00

additional parameters may be introduced into the model. The addition of such parameters is helpful since this increases the "flexibility" of the model to fit the data, but usually has the unfortunate consequences of (i) decreasing the power of model-based statistical tests and (ii) increasing the difficulty of developing simple, insightful models with small numbers of interpretable parameters.

Recently, methods of hypothesis testing (White, 1982, 1989, in press) and model selection (Vuong, 1989) have been developed which are applicable in cases where a given statistical model is misspecified in particular ways. Such methods provide a mechanism for testing statistical hypotheses directly without requiring the assumption that the proposed statistical model is correctly specified with respect to the data generating process. Thus, the researcher can "separate" the model building and evaluation problem from the hypothesis testing problem. Another way to think about this approach to statistical inference is that issues of "model validity" and "model reliability" are separated. In the classical hypothesis testing framework, reliable estimates of the sampling variance of a model parameter cannot be obtained in the presence of model misspecification. Hypothesis testing in a model misspecification framework, however, permits reliable estimates of the sampling variance of a model parameter to be obtained even if the model itself is not completely correct! In fact, if the experimenter has two competing models (each of which is misspecified), Vuong's (1989) statistical framework may be used to decide which of the two competing models is most consistent with the data.

One important limitation of these methods which should be kept in mind is that these methods are based upon "large sample" approximations which are only valid for large data sets. The question of "how large," however, is an empirical question which is discussed in greater detail later by considering alternative "checks" on the asymptotic approximations using Efron's (1982) "boot-strap" methodology. Another important limitation of the proposed model misspecification framework is that the process which generates the data must satisfy certain mild restrictions. For example, both the proposed statistical model and the data generating model are assumed to generate independent and identically distributed (i.i.d.) observations from the same sample space (see White, in press, for less restrictive conditions). The proposed probability model and data generating model must also satisfy additional technical restrictions which usually are satisfied in practice. A third important limitation of the model misspecification framework is that it only addresses the "reliability" issue regarding data analysis, and it cannot be assumed to apply to issues of model "validity." That is, the problem of developing good parsimonious mathematical models which fit the data certainly does not disappear within the model misspecification framework. Rather, the model misspecification framework provides a mechanism for model improvement and data analysis even when the "best" model is not completely accurate in all its details.

This article is organized into two major sections. The purpose of the first part of this article is to select, review, and discuss issues associated with evaluating probability models and constructing asymptotic statistical tests within a framework of model misspecification. In the second part of this article, a parallel distributed processing (PDP) model for analyzing categorical time series data is proposed. A theorem is then stated and proved concerning the uniqueness of the parameter estimates of the PDP model. The PDP model is then used to examine temporal regularities in human free-recall data of short stories (Golden *et al.*, 1993) in order to show (i) explicitly how Vuong's and White's asymptotic statistical theories may be applied to derive new statistical tests for the proposed PDP model and (ii) evaluate the asymptotic statistical theory proposed by Vuong and White with respect to both boot-strap simulations and the classical asymptotic statistical theory.

A REVIEW OF THE MISSPECIFIED MODEL FRAMEWORK

White's (in press) book (also see White, 1982, 1989) provides a comprehensive introduction to the literature of making correct statistical inferences in the presence of model misspecification. To simplify the presentation of the important ideas behind a model misspecification framework approach, the discussion is limited to probability mass functions defined on finite sample spaces. The generalization to the case of probability density functions is straightforward and is thoroughly discussed by White (1982, 1989, in press). White (in press) also shows to relax *considerably* many of the assumptions made for expository reasons in this review.

Hypothesis Testing in the Presence of Model Misspecification

DEFINITION 1. A probability mass function, p, on a finite sample space Ω whose elements are real-valued vectors satisfies (i) for $x \in \Omega$, $0 \leqslant p(x) \leqslant 1$, and (ii) $\sum_{x \in \Omega} p(x) = 1$.

DEFINITION 2 (Adapted from Vuong (1989)). Let a sample space Ω be a finite set of real-valued vectors. A probability mass model, F_W, is a set of probability mass functions on Ω. In addition, the elements of F_W are indexed by the real-valued parameter vector $w \in W$ so that a given element of W refers to the specific probability mass function $q(\cdot, w) \in F_W$.

Note that unlike the classical approach to statistical inference, *two* types of distinct probability mass functions are defined right from the beginning. The unobservable environmental distribution, p_e, which is the data generating process, and the model probability mass function, $q(\cdot, w)$, which is an element of a specific probability model. If the environmental distribution is contained in the probability model, then this corresponds to the classical case. In general, however, the environmental distribution is not contained in the probability model and this is the common case of using a "wrong" or "misspecified" probability model. When working in a model misspecification framework, it is extremely important to be clear about the distinction between the probability model whose parameters are being estimated and the probability distribution which is generating the observations. For this reason, the concept of model misspecification is now formally defined.

DEFINITION 3. Let F_W be a probability mass model. Let p_e be an environmental probability mass function. Then F_W is correctly specified with respect to p_e if $p_e \in F_W$, and F_W is misspecified (i.e., incorrectly specified) with respect to p_e if $p_e \notin F_W$.

Let $x_1, x_2, \ldots, x_n$ be n i.i.d. observations from p_e. The likelihood of the ith observation, x_i, is $q(x_i; w)$ for model probability mass function $q(\cdot; w)$. Moreover, since the n observations are i.i.d., the likelihood of the set of observations, $x_1, \ldots, x_n$, is given by

$$L_n = \prod_{i=1}^{n} q(x_i; w) \qquad (1)$$

since the observations are assumed to be statistically independent. In order to maximize L_n with respect to w, it is often convenient to define a monotonically decreasing function of L_n, $\hat{E}_n(w)$, given by

$$\begin{aligned}\hat{E}_n(w) &= -(1/n)\log[L_n]\\ &= -(1/n)\sum_{i=1}^{n}\log[q(x_i; w)]. \qquad (2)\end{aligned}$$

White (1982) shows that $\hat{E}_n$ converges with probability one to a fixed function, E, of w as n becomes sufficiently large under certain regularity conditions. In particular, the function E is called the Kullback–Leibler (1951) information criterion (KLIC), and is given by

$$E(w) = -\sum_{x \in \Omega} p_e(x)\log[q(x; w)]. \qquad (3)$$

In this article, (2) is referred to as the *sample loss* function, while (3) is referred to as the *true loss* function.

In classical maximum likelihood estimation, the maximum likelihood estimates (MLEs) are those parameters which maximize L_n in (1) or equivalently minimize $\hat{E}_n$ in (2). Maximum likelihood estimates can be shown (under fairly general conditions) to be asymptotically consistent and efficient (Manoukian, 1986; Van Trees, 1968). Moreover, it can be shown that maximum likelihood estimates converge to maximum a posteriori (MAP) (i.e., most probable) estimates under fairly general conditions (Van Trees, 1986).

If a probability mass model F_W is misspecified with respect to an environmental probability mass function, p_e, then the concept of a "true" parameter vector w has no meaning. It is therefore convenient to define the goal of the statistical inference process so that a parameter vector w is sought such that an appropriate measure of the distance between the misspecified probability mass model F_W and the environmental probability mass function p_e is minimized. In addition, the distance measure should be "automatically" consistent with the classical theory so that if in fact the probability mass model F_W is correctly specified with respect to p_e, then the global minima of the distance measure should yield maximum likelihood estimates.

Given these considerations, it seems natural to consider the KLIC function in (3) as an appropriate measure of the distance between the environmental distribution, p_e, and the probability model F_W. Following Kullback and Leibler (1951), the following points must be made about the KLIC as a distance measure. First, the KLIC is not a metric on probability mass functions since it is not symmetric with respect to the environmental and model probability distributions and does not satisfy the triangle inequality. Second, the KLIC distance measure in (3) is never directly observable since p_e is never directly observable. The third and most relevant property is that if the environmental probability distribution, p_e, and probability model, $q(\cdot, w)$, are equivalent, then the KLIC distance between these two distributions obtains its minimum value (Kullback & Leibler, 1951). On the other hand, the KLIC distance function has the following very important property. If the model is correctly specified, minimizing the sample KLIC distance measure in (2) is equivalent to maximizing the log likelihood function (i.e., classical maximum likelihood estimation).

In classical maximum likelihood estimation, it can be shown that the maximum likelihood estimates are normally distributed about the true parameter values with covariance matrix equal to the inverse of the Fisher information matrix (Manoukian, 1986; Van Trees, 1986). White's (1982, 1989, in press) theorems provide a natural generalization of these results to the case where the true parameter values do not exist! In particular, White demonstrates that what he defines as *quasi-maximum likelihood estimates* (i.e., parameter estimates that obtain a global minimum of the sample loss function $\hat{E}_n$) are both consistent (also see Huber, 1967) and asymptotically normally distributed with respect to the global minimum of the KLIC, E, with a covariance matrix which is closely related (but not identical) to the inverse of the Fisher information matrix. It is also important to note that if the probability model happens to be correctly specified, then White's results reduce to the classical maximum likelihood estimation case.

The following two theorems by White (1989; also see White, 1982, in press) summarize these important theoretical results and also provide an explicit procedure for estimating the asymptotic distribution of quasi-maximum likelihood estimates. Again, simplified versions for finite sample spaces of the relevant theorems of White are presented for expository reasons. Generalizations of these results to other important cases such as continuous-valued random variables (White, 1982, 1989) or observations which are not independently distributed (White, 1982, 1989, in press) are available.

THEOREM 1 (White, 1989, Theorem 1). *Let the observations, $\{x_1, ..., x_n\}$ with $x_i \in \mathscr{R}^d$, be i.i.d. according to an environmental probability mass function, p_e, defined on a finite sample space Ω. Let*

$$F_W = \{q(\cdot, w) : \Omega \to \mathscr{R}, w \in W\} \tag{4}$$

be a probability mass model where W is a compact subset of a finite-dimensional Euclidean space and $q(\cdot, w)$ is a probability mass function on Ω for each $w \in W$. Let $\hat{E}_w$ and $E(w)$ be defined respectively as in (2) and (3) with respect to F_w and p_e. Let $\hat{w}_n$ be a strict global minimum of $\hat{E}_n(w)$ on W. Then for each $n = 1, 2, ..., \hat{w}_n$ exists and has the property that $\hat{w}_n \to \Gamma^*$ with probability one, where Γ^* is the set of global minima of $E(w)$ on W.

This theorem (White (1989, Theorem 1)) states that given an environmental probability distribution, p_e, which generates a set of independently and identically distributed vectors $\{x_1, ..., x_n\}$, one can construct a sample loss function, $\hat{E}(w)$, which gradually approximates the true loss function, $E(w)$, more and more accurately with high probability as the number of observations, n, becomes large. Moreover, White's (1989) Theorem 1 says that any strict local minimum of the sample loss function converges with probability one to a set of strict local minima of the true loss function as the number of observations becomes large. Note that the concept of "converge with probability one to a set" means that either (i) the stochastic process will converge to a member of the set or (ii) the stochastic process may oscillate among the members of the set. It is also important to note that although $q(\cdot; w)$ can be misspecified with respect to probability model F_W, Eq. (4) requires that F_W be defined on the same sample space as p_e.

The next theorem (a special case of White's (1989) Theorem 2; also see White, 1982, in press) shows how the asymptotic distribution of the parameter estimates may be derived from a differentiable probability model.

THEOREM 2 (White, 1989, Theorem 2). *Let the observations $\{x_1, ..., x_n\}$ with $x_i \in \mathscr{R}^d$ be i.i.d. according to an environmental probability mass function, p_e, defined on a finite sample space Ω. Let*

$$F_W = \{q(\cdot; w) : \Omega \to \mathscr{R}, w \in W\} \tag{5}$$

be a probability mass model where W is a compact subset of a finite dimensional Euclidean space and $q(\cdot; w)$ is a probability mass function on Ω for each $w \in W$. Assume that $\hat{w}_n$ converges with probability one as n increases to an isolated strict global minimum in W, w^*, of the KLIC function in (3) with respect to $q(\cdot; w) \in F_W$ and p_e. In addition, let $q(x; \cdot)$ for each $x \in \Omega$ have continuous first and second partial derivatives on W. Also assume that

$$A^* = \sum_{x \in \Omega} p_e(x) \nabla_w^2 \log[q(x; w^*)]$$

and

$$B^* = \sum_{x \in \Omega} p_e(x) \nabla_w \log[q(x; w^*)] \nabla_w \log[q(x; w^*)]^T$$

are non-singular matrices, where $\nabla_w \log[q(x; w^*)]$ and $\nabla_w^2 \log[q(x; w^*)]$ are the first and second derivatives of $\log[q(x; \cdot)]$ evaluated at w^* and a^T indicates the transpose of vector a.

Then as n increases, $\sqrt{n}(\hat{w}_n - w^*)$ converges in distribution to a multivariate normal distribution with zero mean and covariance matrix C^*, where

$$C^* = A^{*-1} B^* A^{*-1}. \tag{6}$$

In addition, as n increases,

$$\hat{C}_n = \hat{A}_n^{-1} \hat{B}_n \hat{A}_n^{-1} \tag{7}$$

converges with probability one to C^*, where

$$\hat{A}_n = n^{-1} \sum_{i=1}^{n} \nabla_w^2 \log[q(x_i; \hat{w}_n)] \tag{8}$$

and

$$\hat{B}_n = n^{-1} \sum_{i=1}^{n} \nabla_w \log[q(x_i; \hat{w}_n)] \nabla_w \log[q(x_i; \hat{w}_n)]^T. \tag{9}$$

This theorem says that given an environmental probability distribution which generates a set of independently and identically distributed vectors $\{x_1, ..., x_n\}$, then one can construct a sample loss function, $\hat{E}_n(w)$, as before. Now assume that a strict local minimum, $\hat{w}_n$ of the sample loss function is in fact converging to a specific strict local minimum, w^*, of the true loss function, $E(w^*)$, as the number of observations, n, becomes large. Given this critical assumption and assuming that n is sufficiently large, then this theorem shows that *the asymptotic distribution of $\hat{w}_n$ is Gaussian with mean w^* and covariance matrix C^*/n.* Note that C^* is never directly observable, but White shows that the estimator $\hat{C}_n$ in (7) converges with probability one to the true value C^*. Also note that according to the classical theory of maximum likelihood estimation (e.g. Van Trees, 1986; Wilks, 1962), if the model is correctly specified then the Fisher information matrix equality ($\hat{A}_n = -\hat{B}_n$) holds. Using this equality in (7), we obtain $\hat{C}_n = -\hat{A}_n^{-1} = \hat{B}_n^{-1}$. Thus, White's (1982, 1989) asymptotic statistical theory reduces to the well-known result that the maximum likelihood estimates are asymptotically distributed about their true values with covariance matrix $-\hat{A}_n^{-1} = \hat{B}_n^{-1}$ (Riefer & Batchelder, 1988; Van Trees, 1968; Wilks, 1962).

The first derivative of the sample loss function (i.e., the gradient) should be empirically checked to assess whether the parameter estimates, $\hat{w}_n$, are sufficiently close to their true values, w^*. If $-A^*$ is positive definite, then w^* is a strict local minimum and "locally unique." If B^* is also non-singular, then the covariance matrix C^* is non-singular as well. The non-singularity of C^* is a necessary condition for hypothesis-testing. Note that the non-singularity of A^* and B^* can be empirically checked by examining the rank of $\hat{A}_n$ and $\hat{B}_n$. Finally, it is helpful to note that if the number of parameters of the model is greater than the number of observations, then $\hat{B}_n$ will always be singular.

To understand why White's (1989) Theorem 2 works, a heuristic proof of White's (1989) Theorem 2 is now presented following the discussion in White (in press, Chap. 6). The first step is to do a mean value (Taylor series) expansion of the gradient of the sample loss function about w^* and evaluate the gradient of the sample loss function at $\hat{w}_n$. This can be done since it is assumed that $\hat{w}_n$ is sufficiently close to w^*, and the gradient of the sample loss function has continuous first derivatives. Let $\Delta_n = |\hat{w}_n - w^*|$, then

$$\nabla_w \hat{E}_n(\hat{w}_n) = \nabla_w \hat{E}_n(w^*) + [\nabla_w^2 \hat{E}_n(w^*)](\hat{w}_n - w^*) + O(\Delta_n^2),$$

and since $\nabla_w \hat{E}_n(\hat{w}_n)$ vanishes,

$$\nabla_w \hat{E}_n(w^*) = -[\nabla_w^2 \hat{E}_n(w^*)](\hat{w}_n - w^*) + O(\Delta_n^2).$$

Also since $-\nabla_w^2 \hat{E}_n(w^*) \to A^*$ with probability one by the strong law of large numbers,

$$\nabla_w \hat{E}_n(w^*) = A^*(\hat{w}_n - w^*) + O(\Delta_n^2),$$

with probability one for n sufficiently large. Note that A^* is non-singular so the inverse of A^* always exists. Now, premultiply each side by $[A^*]^{-1}$ to obtain

$$\hat{w}_n - w^* = [A^*]^{-1} \nabla_w \hat{E}_n(w^*) + O(\Delta_n^2), \tag{10}$$

which may be written as

$$\hat{w}_n - w^* = -[A^*]^{-1} n^{-1} \sum_{i=1}^{n} \nabla_w \log[q(x_i; w^*)] + O(\Delta_n^2). \tag{11}$$

Inspection of (11) shows that $\nabla_w \hat{E}_n(w^*)$ has been expressed as the average of independent and identically distributed random vectors, so $\sqrt{n}\,\nabla_w \hat{E}_n(w^*)$ is asymptotically normally distributed with mean vector zero and covariance matrix B^* by the central limit theorem (see Appendix 3 of White, in press, for an overview of such central limit theorems). Also, since $\sqrt{n}\,\nabla_w \hat{E}_n(w^*)$ is asymptotically normally distributed and A^{*-1} is a full rank symmetric linear transformation, $\sqrt{n}(\hat{w}_n - w^*)$ is also asymptotically normally distributed with mean vector zero and covariance matrix $C^* = A^{*-1} B^* A^{*-1}$ by (10).

Model Selection in the Presence of Model Misspecification

This section deals with the problem of deciding which of two probability models is most consistent with a particular set of data. Currently most researchers in the field of mathematical psychology (e.g., Allison & Liker, 1982; Bentler, 1980; Falmagne *et al.*, 1990; Lord, 1980) have used Wilk's (1938) generalized likelihood ratio test (GLRT) for making decisions of this type.

To review the GLRT, the following notation is introduced following Vuong (1989). Let F_W and G_Y be two probability models. If $F_W \subset G_Y$, then we say that F_W is *fully nested* within G_Y and refer to G_Y as the *full model.* In order to correctly use the GLRT to compare two alternative probability models it is absolutely necessary that (i) one probability model be fully nested within the other probability model, and (ii) the full probability model must be correctly specified with respect to the environmental probability distribution (i.e., the data generating process) (e.g., White, 1982; Voung, 1989). In general, it is unrealistic to expect that assumption (ii) *will* be satisfied for substantiative statistical models of complex psychological phenomena containing small numbers of interpretable parameters. In addition, assumption (i) is rather restrictive since one could imagine many situations where the experimental psychologist would want to decide which of two alternative *non-nested* probabilistic models "best fits" a specific data sample.

Vuong (1989) has introduced a general asymptotic statistical theory for probability model selection which allows one to compare two alternative probability models which (i) are not necessarily nested and (ii) may be misspecified with respect to the environmental probability distribution. Moreover, Vuong (1989) shows how his general asymptotic statistical theory for probability model selection is formally equivalent to GLRT in the special case where probability distribution F_W is fully nested in probability distribution G_Y, *and* G_Y is correctly specified.

The basic idea behind Vuong's (1989) theory is to compute the KLIC distance of each of the two probability models with respect to the environmental probability distribution. The model whose KLIC distance to the environmental distribution is shortest is considered to be the better model. Or in other words, the likelihood of the data is computed using probability model F_W and then computed using probability model G_Y. The model which makes the data most likely is then chosen.

More formally, let $q_f(\cdot; w) \in F_W$ and $q_g(\cdot; y) \in G_Y$ be probability mass (or density) functions on a particular sample space. Let p_e be the environmental probability mass (or density) function which generates the observations $\{x_1, ..., x_n\}$. Let w^* be a strict local minimum of the KLIC, $E^f(w)$, between $q_f(\cdot; w)$ and p_e. Let y^* be a strict local minimum of the KLIC, $E^g(y)$, between $q_g(\cdot; y)$ and p_e. Now following Wilk's (1938) generalized likelihood ratio test and Vuong's (1989) statistical theory, form the log likelihood ratio

$$D = E^g(y^*) - E^f(w^*) \tag{12}$$

and choose model F_W if $D > 0$ and model G_Y if $D < 0$. In practice, D is not directly observable so the asymptotic distribution of an appropriate estimate of D is required for hypothesis testing purposes.

Note that Vuong's (1989) model selection theory is still applicable if the parameter vectors are globally identifiable (i.e., w^* and y^* are merely strict local minima). In this "locally identifiable" case, however, the set F_W contains only probability mass (or density) functions of the form $q_f(\cdot, w)$ such that w is in a sufficiently small neighborhood of w^*. Similarly, the set G_Y contains only probability mass (or density) functions of the form $q_g(\cdot, y)$ such that y is in a sufficiently small neighborhood of y^*. From a practical perspective, this means that the *algorithm* used to search for w^* and y^* is implicitly involved in the model identification process. This issue is relevant to certain types of connectionist models, such as the back-propagation algorithm (Rumelhart *et al.*, 1986), which are usually not globally identifiable.

The goal of this section is to informally review Vuong's (1989) asymptotic statistical theory. Readers interested in additional details are urged to consult Vuong (1989). Towards this end, this section of the article is divided into three parts. In the first part, Wilk's (1938) GLRT is briefly and informally reviewed. In the second part, Vuong's (1989) statistical test for comparing two probability models which are known to be strictly non-nested is briefly and informally reviewed. The probability models are not required to be correctly specified in this case. The third part of this section reviews Vuong's (1989) general asymptotic statistical theory which is applicable to situations where (i) the probability models are not necessarily fully nested and (ii) it is not required that either of the probability models is correctly specified.

Finally, the following notation is used throughout the next several sections. Let $\hat{y}_n \to y^*$ and $\hat{w}_n \to w^*$ as $n \to \infty$ with probability one. The estimated KLIC distance of $q_f(\cdot; w^*)$ to p_e is given by

$$\hat{E}^f_n(\hat{w}_n) = -n^{-1} \sum_{i=1}^{n} \log[q_f(x_i; \hat{w}_n)] \tag{13}$$

and the estimated KLIC distance of $q_g(\cdot; y^*)$ to p_e is given by

$$\hat{E}^g_n(\hat{y}_n) = -n^{-1} \sum_{i=1}^{n} \log[q_g(x_i; \hat{y}_n)]. \tag{14}$$

The estimated log likelihood ratio is given by

$$\hat{D}_n = \hat{E}^g_n(\hat{y}_n) - \hat{E}^f_n(\hat{w}_n).$$

The Fully Nested and Correctly Specified Case (GLRT). When the two probability models F_W and G_Y have the properties that (i) $F_W \subset G_Y$ and (ii) G_Y is correctly specified, then the GLRT is applicable given that some additional "regularity" conditions are satisfied (Wilks, 1938; see Vuong, 1989, for a presentation using modern notation). These additional "regularity" conditions essentially assume that certain expectations (taken with respect to the environmental probability distribution) of the probability distributions associated with F_W and G_Y exist and are sufficiently smooth. According to Wilk's (1938) generalized likelihood ratio test (also Vuong, 1989), $-2n\hat{D}_n$ has an asymptotic chi-square distribution with $p - q$ degrees of freedom where p is the dimensionality of the full model parameter vector, y, and q is the dimensionality of the reduced model parameter vector w.

To use the GLRT, one therefore follows the following procedure. First, compute quasi-maximum likelihood estimates $\hat{y}_n$ and $\hat{w}_n$ for each of the two probability models G_Y and F_W. Second, check that the full model G_Y provides a good fit to the data. Third, use a standard cumulative chi-square distribution table to compute the critical value, $\chi^2_\alpha(p-q)$, which has the property that a chi-square random variable with $p - q$ degrees of freedom will exceed $\chi^2_\alpha(p-q)$ with probability (significance level) α. Fourth, reject the null hypothesis that F_W and G_Y are equally distant from p_e if the statistic

$$\hat{S}_n = -2n\hat{D}_n$$

exceeds the critical value $\chi^2_\alpha(p-q)$.

The Strictly Non-nested Case. Now consider the strictly non-nested case where $q_f(\cdot; w^*)$ and $q_g(\cdot; y^*)$ are not identical functions on Ω. In addition, this section allows either F_W or G_Y to be misspecified with respect to the environmental probability distribution. Again, specific regularity conditions concerned with the existence and smoothness of certain types of expectations of the probability functions in F_W and G_Y with respect to the environmental distribution are assumed to be satisfied.

Define

$$\hat{\sigma}^2_{v_n} = (1/n) \sum_{i=1}^{n} (\log[q_f(x_i \mid \hat{w}_n)] - \log[q_g(x_i \mid \hat{y}_n)])^2. \tag{16}$$

Also let $\hat{D}_n$ be defined as in (15). For the strictly non-nested case, Vuong (1989) proved that as n increases, the model selection statistic

$$\hat{V}_n = \hat{D}_n/[\hat{\sigma}_{v_n}/n^{1/2}] \tag{17}$$

converges in distribution to a normally distributed random variable with mean zero and variance one in the case that F_W and G_Y are equally distant from the environmental distribution, p_e. The derivation of the asymptotic distribution of $\hat{V}_n$ is based upon the idea that (15) is the average of n independent and identically distributed random variables. Thus, $\hat{D}_n$ has a Gaussian distribution with variance as in (16) by the central limit theorem (see White, in press, for a review of such theorems).

Thus to decide which of two strictly non-nested probability models "best fits" the data, the following computationally simple hypothesis-testing procedure may be used. First, compute quasi-maximum likelihood estimates $\hat{y}_n$ and $\hat{w}_n$ for each of the two probability models G_Y and F_W. Second, let Z_α be defined such that the probability that a normally distributed random variable with mean zero and variance one has a magnitude greater than Z_α is α. Third, compute $\hat{V}_n$ as in (17). Then decide F_W and G_Y are equally distant from p_e (i.e., $H_0: D=0$) if $|\hat{V}_n| < Z_\alpha$. Decide F_W is closer to p_e than G_Y (i.e., $H_F: D>0$) if $\hat{V}_n > Z_\alpha$. And finally, decide G_Y is closer to p_e than F_W (i.e., $H_G: D<0$) if $\hat{V}_n < -Z_\alpha$.

The General Case. This section considers the most general case where the probability models F_W and G_Y may be either (i) correctly specified or misspecified, and (ii) the nesting relationship between the two probability models is unknown. Again certain regularity conditions concerned with the existence and smoothness of certain expectations of the probability functions in F_W and G_Y are assumed (see Vuong, 1989, for additional details).

Vuong's (1989) hypothesis-testing procedure for the general case is organized into two distinct stages. In the first stage, a statistical test called the "variance test" (reviewed in Appendix 3) is done to decide if the two probability models, F_W and G_Y, are strictly non-nested (i.e., decide if $H_0: D=0$). If the null hypothesis of the variance test is rejected, then conclude that F_W and G_Y are strictly non-nested and proceed to the second stage of the analysis where the statistical test for strictly non-nested models is then done. The second stage of the analysis is then used to decide whether (i) model F_W is better than model G_Y, (ii) model G_Y is better than F_W, or (iii) the information in the data set is not sufficient for deciding which model is better.

Two important additional comments also need to be made. First, Vuong (1989) shows that if the significance level for the variance test is α and the significance level for the strictly non-nested test is α, then the significance level for the entire two-stage procedure is no larger than α. And second, it should be emphasized that the variance test is a complex statistical test which requires more computational resources than the second stage of Vuong's (1989) analysis. Fortunately, the variance test can be replaced with any alternative statistical test or analysis to decide if $q_f(\cdot; w^*)$ and $q_g(\cdot; y^*)$ are equivalent (Vuong, 1989, Lemma 4.1).

Validity of the Asymptotic Approximations

Both White's (1982, 1989, in press) theory of covariance matrix estimation and Vuong's (1989) theory of model selection are based upon approximations which are only valid for sufficiently large numbers of observations. In this section, a computer simulation methodology is introduced for empirically checking the validity of these proposed asymptotic approximations for specific probability models and specific data sets.

The typical approach to checking the validity of asymptotic approximations is to choose some "true" parameters for the assumed probability distribution, generate sample observations from that distribution, and select one sample of size n from the generated data. Then, for that sample of size n, the parameters and their asymptotic variance are estimated using either the asymptotic approximations proposed by White (1982, 1989, in press) or Vuong (1989). From this information, the parameter estimates and their standard errors can be computed from the data sample using classical asymptotic statistical theory and compared to the original "true" parameters.

This typical Monte Carlo approach will not work, however, when one assumes that the parameterized probability model may be misspecified because the typical Monte Carlo approach automatically guarantees that the probability model is correctly specified! An alternative to the typical Monte Carlo approach is the boot-strap approach proposed by Efron (1982). The basic idea of the boot strap approach is to use actual data, $x_1, ..., x_n$, to empirically estimate the environmental probability distribution by assuming that the probability of x_j $(j = 1, ..., n)$ is equal to $1/n$. For example, if the data consisted of the observations 2, 3, 3, 3, then the probability of an observation with value 3 would be estimated to be

$$\tfrac{1}{4}+\tfrac{1}{4}+\tfrac{1}{4}=\tfrac{3}{4},$$

since observations x_2, x_3, and x_4 have the same value, 3. More generally, suppose the environmental probability distribution has the parametric form $p(x_l) = p_l$ $(l = 1, ..., M)$ where the free parameters p_l satisfy (i) $0 \leqslant p_l \leqslant 1$ and (ii) $\sum_{l=1}^{M} p_l = 1$. Let x^l denote the lth value of the random variable of interest. Then it is not difficult to show that the quasi-maximum likelihood estimate, $\hat{p}_l$, of p_l is given by the relative frequency of occurrence of x^l in the data sample.

Given an estimated environmental probability distribution, one then generates sample observations and selects K samples of size n from the generated data. Then, with respect to the ith sample $(i = 1, ..., K)$ of size n, the quasi-maximum likelihood estimates $\hat{w}_n^i$ are computed. Thus, the parameters associated with the ith sample are denoted by $\hat{w}_n^i$. Let $\hat{w}_n$ be the *boot-strap estimate* of w. Let $\hat{C}_n$ be the boot-strap estimate of the covariance matrix of w (i.e., an estimate of C^* in (6)).

Then, as $K \to \infty$,

$$(1/K) \sum_{i=1}^{K} \hat{w}_n^i \to \hat{w}_n$$

and

$$(1/K) \sum_{i=1}^{K} [\hat{w}_n^i - \hat{w}_n][\hat{w}_n^i - \hat{w}_n]^T \to \hat{C}_n.$$

Note that the boot-strap method is also based upon asymptotic (large sample) approximations since the estimated environmental probability distribution only approaches the actual environmental probabily distribution as the number of observations, n, in the original data set becomes large. The boot-strap estimates can then be compared to the parameter and covariance matrix estimates derived by White (1982, 1989, in press) for the original sample of size n. An agreement between the two very different asymptotic approximations for estimating the covariance matrix of the parameter estimates would provide some reassurance that the sample size, n, is sufficiently large. Efron (1982) provides a good review of the boot-strap algorithm. The fundamental problem with the boot-strap method is that it is very computationally intensive.

The Parameter Estimation Problem

The practical application of the asymptotic model misspecification theory reviewed here depends upon finding a strict local minimum, $\hat{w}_n$, of the sample KLIC function. Numerical methods for finding such minima of the sample KLIC function are well-known in the optimization and neutral network literature (e.g., Golden, 1988a, 1988b, 1988c; Luenberger, 1984; White, 1989).

For example, a gradient descent type scheme may be used which has the form

$$\hat{w}_n^{i+1} = \hat{w}_n^i - \gamma_i \nabla \hat{E}_n(\hat{w}_n^i), \qquad (18)$$

where $\hat{w}_n^0$ is an initial guess about the parameter vector values and $\hat{w}_n^i$ is the estimate of the parameter values at iteration i of the algorithm. The sequence of step constants, $\gamma_1, \gamma_2, ...$, is determined at each step of the algorithm so that γ_i is a global minimum of $\hat{E}_n(\hat{w}_n^i)$. The quantity $\nabla \hat{E}_n(\hat{w}_n^i)$ is the derivative of the loss function $\hat{E}_n(w)$ with respect to the parameter vector w evaluated at the current estimate, $\hat{w}_n^i$, of the parameter vector w. Of course the problem with any numerical scheme is that there will always be an intrinsic error associated with the parameter estimation process which is a source of error in the resulting parameter estimates. It should be clear at this point that this intrinsic numerical error due to computational limitations in exactly computing $\hat{w}_n$ has been assumed to be significantly smaller in magnitude than the sampling error (i.e., the error associated with minimizing the sample KLIC function rather than the true KLIC function).

Correlations among Parameter Estimates

One important virtue of methods such as balanced analysis of variance (ANOVA) is that the models are constructed so that the factors are orthogonal and thus the correlations among the parameters are zero. A similar issue arises in multiple linear regression analysis where the problem of multicollinearity is considered (Montgomery & Peck, 1982). All of the above issues can be reformulated in a more general setting by considering the following key question: To what extent are the parameters of a probability model correlated? Note that the issue of parameter correlation is relevant for correctly specified as well as misspecified models.

A model with highly correlated parameters will have two undesirable properties. First, it is much more difficult to interpret the effects of individual parameters on the model's behavior. Second, highly correlated parameter estimates imply that the "true" Hessian of the error function may be ill-conditioned in the vicinity of the parameter estimates which is a violation of one of the key critical assumptions of White's (1982, 1989, in press) theory, Vuong's (1989) theory, classical theories based upon using the Fisher information matrix for estimating the covariance matrix, and Wilk's (1938) generalized likelihood ratio test.

For these reasons, it is useful to examine the correlations among the model parameter estimates. This calculation is easily done by inspection of the asymptotic covariance matrix $\hat{C}_n$ which is computed using (7). Let $\hat{w}_n(i)$ be the ith element of the strict local minimum, $\hat{w}_n$, of the sample KLIC loss function such that $\hat{w}_n$ approaches the strict local minimum, w^*, of the true KLIC loss function as n increases. Let $\hat{c}_{ij}$ be the ijth element of the asymptotic covariance matrix of the parameter estimates $\hat{C}_n$. Then an estimate, $\hat{r}_n(i, j)$, of the correlation between $\hat{w}_n(i)$ and $\hat{w}_n(j)$ is given by

$$\hat{r}_n(i, j) = \frac{\hat{c}_{ij}}{\sqrt{\hat{c}_{ii}\hat{c}_{jj}}}.$$

A test of statistal significance of the quantity $\hat{r}_n(i, j)$ can be constructed by estimating the asymptotic variance of $\hat{r}_n(i, j)$ using boot-strap techniques.

Within-Groups Hypothesis Testing

Define a *within-groups* comparison to be based upon a single set of observations (measurements), $\{x_1, ..., x_n\}$. If $\hat{w}_n$ is the parameter estimate of the unobservable parameter w^*, then a Wald test as suggested by White (1982, in press) may be constructed to decide whether to reject the null hypothesis,

$$H_0 : Sw^* = 0,$$

where S is a constant *selection* matrix, and 0 is a vector of zeros. Let the rank of the selection matrix S be r. The quantity

$$\mathcal{W}_n = n\hat{w}_n^T S^T [S\hat{C}_n S^T]^{-1} S\hat{w}_n \overset{A}{\sim} \chi_r^2, \tag{19}$$

where χ_r^2 is a chi-square random variable with r degrees of freedom, and the notation $\mathcal{W}_n \overset{A}{\sim} \chi_r^2$ indicates that $\mathcal{W}_n$ is asymptotically distributed according to a chi-square distribution with r degrees of freedom.

The mechanics of the test are thus straightforward. Compute the statistic $\mathcal{W}_n$. Then, if $\mathcal{W}_n > \chi_\alpha^2(r)$, reject the null hypothesis H_0: $Sw^* = 0$. Note that if w^* is not a *unique strict global minimum* and is merely a strict local minimum, then w^* is only locally identifiable. This means, as previously noted in the discussion on model selection statistical tests, that the parameter estimates are *substantive only if characteristics of the parameter estimation procedure are taken into account as well.*

If the null hypothesis is not rejected but the experimenter believes that the source of the problem is due to a choice of n which was too small, it is possible to estimate the sample size, $\hat{n}'$, necessary for a future replication of the original experiment. A $\mathcal{W}_{\hat{n}'}$ is desired so that

$$\mathcal{W}_{\hat{n}'} = \hat{n}'\hat{w}_{\hat{n}'}^T S^T [S\hat{C}_{\hat{n}'} S^T]^{-1} S\hat{w}_{\hat{n}'} > \chi_\alpha^2(r) \tag{20}$$

at the desired significance level, α, and where r is the rank of the selection matrix S. Using $\hat{C}_n$ and $\hat{w}_n$ as estimates of $\hat{C}_{\hat{n}'}$ and $\hat{w}_{\hat{n}'}$ respectively, and solving (20) for $\hat{n}'$,

$$\hat{n}' > n[\chi_\alpha^2(r)/\mathcal{W}_n].$$

Between-Groups Hypothesis Testing

A *between-groups* comparison is based upon two or more sets of independent measurements. For example, these measurements could be obtained by measuring the performance of one group of subjects at two different points in time. Or alternatively, these measurements could be obtained by measuring the performance of two groups of subjects at the same point in time. That is, the terminology "within-groups" and "between-groups" which is used here should not be confused with the terminology "within-subjects" and "between-subjects" since the concept of a "subject factor" is not relevant here (at least not in the usual sense).

For a between-groups comparison, the asymptotic distribution of the parameter estimates of the first data set is computed using model F_W. Let the estimated mean and the covariance matrix of these parameter estimates be denoted as $\hat{m}_1$ and $\hat{c}_1$, respectively. The asymptotic distribution of the parameter estimates of the second data set is then computed using exactly the same probabilistic model F_W. Let the estimated mean and the covariance matrix of the parameter estimates derived from the second data set be denoted as $\hat{m}_2$ and $\hat{c}_2$, respectively.

Now since the estimates $\hat{m}_1$ and $\hat{m}_2$ from the two groups are statistically independent and normally distributed, their joint distribution is normally distributed with mean $\hat{m}$ and covariance matrix $\hat{c}$ given by the formulas $\hat{m} = [\hat{m}_1, \hat{m}_2]$ and

$$\hat{c} = \begin{bmatrix} \hat{c}_1 & 0 \\ 0 & \hat{c}_2 \end{bmatrix}, \tag{21}$$

where 0 is a d-dimensional submatrix of zeros. A large selection matrix, S, may then be constructed for the between-groups hypothesis-testing problem in exactly the same manner as such a matrix was constructed for the within-groups case considered in the previous section. Note that these ideas are easily generalized to the case of G independent groups. In this latter case, $\hat{c}$ will have G submatrices along its main diagonal.

Multiple Wald Tests on the Same Data Set

Typically the experiment-wise significance level (i.e., the probability of a Type I error), α_e, is chosen to be 0.05 or 0.01 in the psychology literature. Note that non-independent planned or post-hoc comparisons are known as non-orthogonal contrasts in the analysis-of-variance literature. Multiple Wald tests on the same data set usually have to be treated as multiple non-independent comparisons for nonlinear statistical models.

If multiple non-independent comparisons using the Wald test need to be made on the same data set, then the experiment-wise probability of a Type I error, α_e, must be properly controlled. Let α_c denote the probability of a Type I error (i.e., the significance level) associated with a specific statistical test on a data set. Let α_e be the probability that for a set of K non-independent comparisons, one or more comparisons resulted in a Type I error. It then follows from the Bonferroni inequality (Manoukian, 1986), p. 4) that

$$\alpha_e \leqslant K\alpha_c, \tag{22}$$

where K is the number of non-independent statistical tests. The Bonferroni inequality suggests that a conservative significance level, α_c, for each individual test may be chosen so that $\alpha_c = \alpha_e/K$ to guarantee that the experiment-wise significance level does not exceed α_e.

A PDP MODEL FOR CATEGORICAL TIME SERIES PATH ANALYSIS

In this section a new PDP model for analyzing data from categorical (nominal) time series is introduced. In addition, it is shown that under certain conditions, which are usually satisfied in practice, the quasi-maximum likelihood estimates of this new model are unique. Explicit formulas for the asymptotic variance of the parameter estimates and methods for model selection using White's and Vuong's asymptotic statistical theory are also derived and used. Finally, an empirical comparison between some aspects of White's (1982, 1989, in press) and Vuong's (1989) asymptotic statistical theories, Efron's (1982) boot-strap approach, and more classical approaches to estimating the variance of the parameter estimates is made.

Consider a complex system which can enter into only one of d states, $f_1, f_2, ..., f_d$, at each instant in time. For example, the states of the system might correspond to a set of d "concepts" in a story. A particular subject's recall of the story from memory would be represented by an ordered sequence of these concepts such as

$$\{f_1, f_3, f_5, f_5, f_2, f_9\}. \tag{23}$$

That is, the "trajectory" in (23) would indicate that concept f_1 was recalled first, then concept f_3 was recalled, then f_5 was recalled twice, then f_2 was recalled, and finally f_9 was recalled. The number of concepts recalled by the subject is in this case equal to six. Thus, the "trajectory length" is equal to six. Note that the trajectory length may be considered to be a random variable as well. Categorical time-series stochastic processes occur in many areas of cognitive science and experimental psychology. For example, Allison and Liker (1982) have used categorical time-series models to study social interactions between individuals. As another example, hidden Markov models of speech recognition (Levinson *et al.*, 1988) are based upon categorical time-series stochastic processes.

A third example of the application of the categorical time-series analysis approach is described by Golden *et al.* (1993). Golden *et al.* (1993) used a categorical time-series analysis model described to analyze the temporal structure in story recall data. Furthermore, the proposed categorical time-series analysis model was shown to be formally equivalent to a special type or highly constrained parallel distributed processing network.

This section is organized in the following manner. First, the basic statistical model for categorical time-series analysis is proposed and the PDP interpretation of the model is briefly noted. Second, a theorem is stated and proved which provides explicit conditions for guaranteeing the uniqueness of the quasi-maximum likelihood estimates of the model (Appendix 2). Third, explicit formulas for hypothesis testing and model selection relating the asymptotic statistical theory developed in the previous section of this article to the model are derived (Appendix 1). And fourth, simulation results of the proposed PDP model are discussed in order to empirically evaluate the applicability of White's and Vuong's asymptotic statistical theories with respect to the proposed PDP model and the data set of Golden *et al.* (1993).

The Probabilistic Model

It is assumed that each *observation* or trajectory corresponds to a point in some sample space Ω. The proposed probability mass model is defined with respect to Ω. The form of the ith observation, x^i, is a d-dimensional by T^i-dimensional matrix of the form

$$x^i = [x^i(1), x^i(2), ..., x^i(T^i)],$$

where T^i is a positive integer and $x^i(1) = K$ where K is a constant vector. The tth column of x^i is a d-dimensional column vector, $x^i(t)$, which can only take on the values $\{u_1, ..., u_d\}$ where u_l is a d-dimensional vector with the lth element equal to one indicating that category label l was observed at time t. The remaining $d-1$ elements of u_l are set to zero. Note that although it is assumed that each observation (trajectory) has a definitive starting point, the *trajectory length* of the ith observation (trajectory) is assumed to be the value, T^i, of a random variable, T.

For example, the seven-dimensional by four-dimensional observation

$$x^3 = [K, u_2, u_1, u_6]$$

indicates that observation or data point number three is identified as the categorical time series where category u_2 is observed after initial condition K, then category u_1 is observed, and finally category u_6 is observed. The value, T^3, of the random variable, T, associated with observation number three has the value of four.

Let $x^i = [x(1) \cdots x(T^i)]$ be the ith observation of N observations. Let $y^i(1)$ be a d-dimensional vector of zeros. Define for $t = 2$ to $t = T^i$,

$$y^i = \Phi(x^i(t-1), y^i(t-1)), \tag{24}$$

where $\Phi\colon \mathscr{R}^d \times \mathscr{R}^d \to \mathscr{R}^d$ has continuous second partial derivatives.

Define the d-dimensional by h-dimensional matrix, $v^i(t)$, as

$$v^i(t) = [w^1 y^i(t) \cdots w^h y^i(t)],$$

where w^l is a d-dimensional matrix of constants which is referred to as a "digraph" matrix.

Let

$$h^i(t) = v^i(t)\alpha, \tag{25}$$

where the jth element of $h^i(t)$ is $h^i_j(t)$ and the lth element of α is α_l. Then, for $t = 2$ to $t = T^i$, define

$$q^i_j(t) = \exp(h^i_j(t)) \Big/ \sum_{k=1}^{d} \exp(h^i_k(t)). \tag{26}$$

Finally, let the prior distribution of T be a Poisson distribution of the form

$$q(T^i) = \lambda^{T^i} \exp(-\lambda)/T^i!, \tag{27}$$

where T^i is the ith positive integer value of random variable T.

Then the probability of trajectory observation x^i is defined as

$$q(x^i \mid x^i(1)) = q(T^i) \prod_{t=2}^{T^i} [x^i(t)^T q^i(t)], \tag{28}$$

where $q^i(t)$ is a d-dimensional vector whose jth element is $q^i_j(t)$.

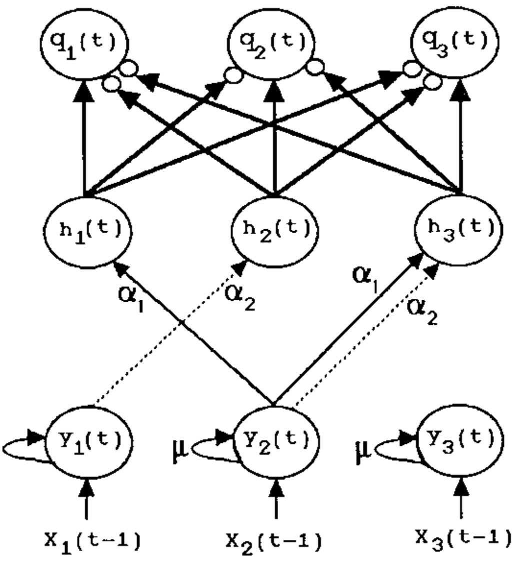

FIG. 1. PDP model interpretation of the categorical time-series analysis probabilistic model. The activation of the ith input unit, $y_i(t)$, is updated as a weighted sum of its past value and the current input to that unit, $x_i(t)$, using the μ constant. Connections from the input units to jth hidden unit, $h_j(t)$, are constrained such that a smaller number of free parameters, $\alpha_1, ..., \alpha_h$ completely specify all connection strengths. In this example with $h = 2$, the solid connections have strength α_1 and the dashed connections have strength α_2. The pattern of connection from the hidden units to the kth probability unit, $q_k(t)$, is not modifiable and effectively implements a forward lateral inhibition network.

Figure 1 shows this probability model is formally equivalent to a special type of highly constrained connectionist network. For expository reasons, consider a Jordan style connectionist network (Jordan, 1992) with an exponentially decaying memory. For networks of this type, the Φ function in (24) can be defined recursively as

$$\begin{aligned} y^i(t) &= \Phi(x^i(t-1), y^i(t-1)) \\ &= x^i(t-1) + \mu y^i(t-1), \end{aligned} \tag{29}$$

where $y^i(1)$ is a vector of zeroes.

The ith activation pattern over the input units at time t is specified by $y^i(t)$ which is updated according to (24) using incoming activation pattern $x^i(t-1)$ and memory buffer $y^i(t-1)$. The lth matrix, w^l, can be interpreted as corresponding to a set of connections from the input to the hidden units in Fig. 1 which are constrained so that all connections in that set have some value α_l. The activation pattern over the hidden units at time t specified by the vector $h^i(t)$. And finally, the mapping from the hidden unit activation pattern $h^i(t)$ to the activation pattern over the output units, $q^i(t)$, specified by (26) can be interpreted as a type of forward lateral inhibition network of fixed (non-modifiable) connections.

Estimation of Model Parameters and Their Asymptotic Variance

The sample loss function, $\hat{E}_n(\alpha, \lambda)$, is derived by substituting (28) into (2) to obtain

$$\hat{E}_n(\alpha, \lambda) = -(1/n) \sum_{i=1}^{n} \log[q(x^i \mid x^i(1))],$$

where $q(x^i \mid x^i(1))$ is defined by Eqs. (24)–(28). The quasi-maximum likelihood estimates $(\hat{\alpha}, \hat{\lambda})$ are the critical points of $\hat{E}_n(\alpha, \lambda)$. Appendix 1 shows that the critical points $(\hat{\alpha}, \hat{\lambda})$ are the solutions to the system of equations

$$d\hat{E}_n(\hat{\alpha}, \hat{\lambda})/d\alpha = 0_h, \tag{30}$$

where

$$\hat{\lambda} = (1/n) \sum_{i=1}^{n} T^i. \tag{31}$$

Appendix 1 also provides an explicit formula for (30) and shows how the asymptotic covariance matrices which are required for White's (1982, 1989, in press) and Vuong's (1989) asymptotic statistical theories may be computed.

Finally, the following theorem regarding the uniqueness of the parameter estimates is proved in Appendix 2.

Theorem 3. *Let $\hat{E}_n(\alpha, \lambda)$ be the Hessian of the sample loss function defined by* (38) *in Appendix* 1. *Let $(\hat{\alpha}, \hat{\lambda})$ be a critical point of $\hat{E}_n(\alpha, \lambda)$ which is computed using* (30) *and* (31). *If the eigenvalues of the Hessian of $\hat{E}_n(\alpha, \lambda)$ evaluated at $(\hat{\alpha}, \hat{\lambda})$ are strictly positive, then $(\hat{\alpha}, \hat{\lambda})$ is the unique quasi-maximum likelihood estimate (i.e., strict global minimum of $\hat{E}_n(\alpha, \lambda)$).*

Empirical Evaluation of the Asymptotic Approximations

In this section, the adult recall data collected by Golden *et al.* (1993) is used to empirically evaluate the validity of the proposed asymptotic approximations with respect to (i) a boot strap Monte Carlo method of estimating the variance of the parameter estimates and (ii) classical asymptotic statistical methods for estimating the variance of the parameter estimates. The model Golden *et al.* (1993) considered was of the form of (28) with a parameter vector of dimension 4 and 24 observations. The covariance matrices for eight different independent data sets (four different texts recalled in an immediate recall and delayed recall condition) were then combined to form a single 32-dimensional covariance matrix as previously discussed.

Using the formulas derived in Appendix 1, the asymptotic covariance matrices $-\hat{A}_n^{-1}$, $\hat{B}_n^{-1}$, and $\hat{C}_n$ were evaluated at the parameter estimates obtained from solving (30) and (31). As previously noted, if the model is correctly specified with respect to the data then White's asymptotic covariance matrix $\hat{C}_n$ should have roughly the same value as the asymptotic covariance matrices $(-\hat{A}_n^{-1}$ and $\hat{B}_n^{-1})$ derived from classical asymptotic statistical theory. On the other hand, if $-\hat{A}_n^{-1}$ differs from $\hat{B}_n^{-1}$, then this signals the presence of model misspecification and it is not correct to use either $-\hat{A}_n^{-1}$ or $\hat{B}_n^{-1}$ as estimates of the asymptotic covariance matrix of the parameter estimates. Rather, in this latter case, one should use White's formula which combines the information in both $\hat{A}_n^{-1}$ and $\hat{B}_n^{-1}$ correctly to yield the estimated asymptotic covariance matrix $\hat{C}_n$.

Using a Sparc 10 workstation, these parameter estimates and their asymptotic variances were obtained within an hour or two of CPU time. Next, boot-strap estimates of the parameter values and their respective covariance matrices were computed. This latter calculation required several days of CPU time. The eigenvalues of the 32-dimensional asympototic covariance matrices were compared with the eigenvalues of the covariance matrices calculated from boot-strap simulations. Rather than plot the eigenvalues directly, the negative natural logarithm of each eigenvalue was plotted in order to compress the eigenspectrum.

As can be seen in Fig. 2, although the eigenspectrum derived from the boot-strap covariance matrix estimates using 20 samples (dashed line with crosses) differ considerably from the other asymptotic estimates, after 300 samples (open circles) the boot-strap covariance matrix estimates converge to the predictions of White's asymptotic

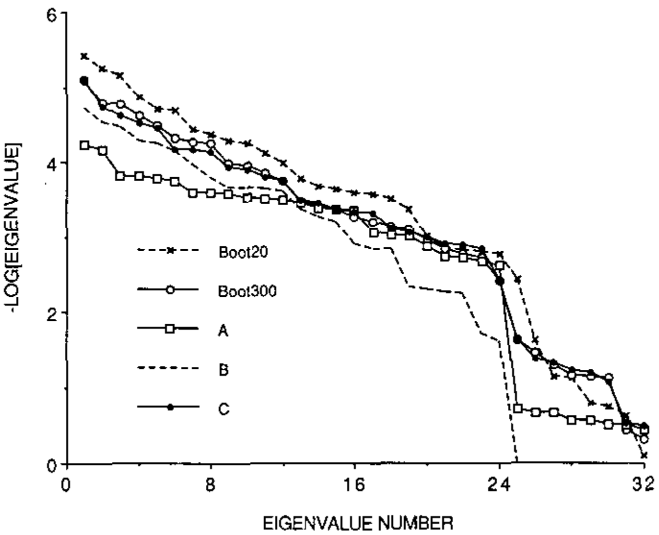

FIG. 2. Comparison of negative logarithm of the eigenspectrums of alternative asymptotic covariance matrix estimates (see text for additional details). Eigenspectrum estimated from White's theory (solid circles) is consistent with boot-strap estimates computed from 300 boot-strap samples (open circles). Also note the two eigenspectrums estimated using classical asymptotic statistics (open squares and dashed lines without crosses) are not only inconsistent with the boot-strap estimates (open circles) but they are inconsistent with each other as well. The eigenspectrum represented by dashed lines with crosses is the boot-strap eigenspectrum computed from only 20 boot-strap samples.

statistical theory (solid circles). Moreover, the observed difference in Fig. 2 between the negative logarithm of the *eigenspectrum of* $-\hat{A}_n^{-1}$ *(open squares) and* $\hat{B}_n^{-1}$ *(dashed lines* without crosses) indicates the failure of the Fisher information matrix equality ($-\hat{A}_n = \hat{B}_n$) and suggests the presence of model misspecification. The presence of model misspecification explains why the asymptotic covariance matrices, $-\hat{A}_n^{-1}$ and $\hat{B}_n^{-1}$, associated with classical asymptotic statistical theory do not agree as effecively with the boot-strap covariance matrix estimates (see Fig. 2) relative to the predictions of White's asymptotic statistical theory.

These simulation results indicate that, at least for the application considered by Golden *et al.* (1993), reliable statistical inferences in the presence of model misspecification can be successfully made using White's statistical theory. These simulation results also indicate that the use of the inverse of the classical Fisher information matrix (i.e., $-\hat{A}_n^{-1}$ or $\hat{B}_n^{-1}$) can lead to wrong statistical inferences in the presence of model misspecification.

GENERAL DISCUSSION

The tools and techniques proposed in this article should prove useful to researchers for at least four specific reasons. First, such methods can provide principled and disciplined guidance to approaching the model selection problem in the early stages of mathematical model building. Second, such methods provide a mechanism for making explicit statistical predictions from a probabilistic psychological theory which may only be partially correct. Third, Vuong's (1989) model selection theory is not limited to comparing nested statistical models as in the Wilk's (1938) generalized likelihood ratio test (also known as the χ^2 test), but can be used to compare models which have equal numbers of parameters. And fourth, such methods provide a mechanism for making correct statistical inferences from correctly specified models which might suffer from trivial flaws.

Indeed, it is possible some researchers who are currently using GLRT are making wrong statistical inferences because the assumed full model is not an entirely accurate representation of the true data generating process. Similarly, it is a common assumption among researchers that the assumed statistical model fits the data well enough that the Fisher information matrix equality is valid. In situations where the parametric model does not contain the data generating process, however, this assumption is not justified. In addition, without this assumption, classical expressions for asymptotic covariance matrices of maximum likelihood estimates result in formulas which are simply incorrect. White (1982, 1989, in press) has provided formulas for asymptotic covariance matrices which are valid even when extreme violations of the Fisher information matrix equality occur (i.e., when the parametric model is not correct).

Like the GLRT and classical asymptotic methods for estimating the covariance matrix of the parameter estimates, the theory presented here is only valid for large samples. If the assumed statistical model is correctly specified, then the validity of the large sample assumptions can be evaluated by simulation experiments designed to examine if the model's "true" parameters (i.e., the parameters which actually generated the test data) can actually be recovered by a "typical" sample size. Unfortunately, the assumption of correct specification is not appropriate in many situations. Indeed, one suspects that if many researchers were to bother to check the validity of (i) large sample assumption and (ii) the correctly specified assumption of the GLRT with respect to their applications, it would not be surprising to this author if many applications of the GLRT were not technically appropriate (i.e., wrong). This paper has suggested a possible solution to the problem of checking the reliability of large sample statistical inferences by exploiting Efron's (1982) boot-strap method. The basic idea of this procedure is to use the data as the basis for an asymptotic non-parametric model of the statistical environment and then evaluate parametric asymptotic approximations with respect to the asymptotic non-parametric model.

The difficulty of deriving and programming the new statistical tests only requires slightly more work than the derivation of statistical tests based upon classical methods. For example, White's (1982, 1989, in press) method for estimating the asymptotic covariance matrix of the parameter estimates is essentially a way of combining both the first and second derivative information derived from the assumed statistical loss function. It is usually not difficult to derive, program, and compute both the first and second derivatives of the loss function. On the other hand, a little more work is required to derive and program the first part of Vuong's (1989) theory. Nevertheless, one can avoid this portion of the theory by either (i) making the possibly inappropriate assumption that the population variance of the model error difference between the two non-nested models is different from zero or (ii) proving that the two models are strictly non-nested. The second part of Vuong's (1989) theory is then computationally trivial: Use the mean and standard error of the model error difference across data samples to decide if the model error difference is significantly different from zero.

Of course the boot-strap calculations are computationally expensive but those calculations are designed to empirically validate the large sample approximations. The computational expense of the boot-strap calculations should not be considered to be a particular problem associated with the theoretical framework described here. Any researcher proposing a new large sample statistical tests should eventually exploit such computer techniques to investigate the validity of the large sample assumptions.

Indeed, if such computations are computationally prohibitive, the statistical theory described in this article is even more applicable since it generalizes classical large sample methods by making fewer assumptions about the fit of the model to the data generating process.

Finally, the key analytical calculations associated with White's (1982, 1989, in press) and Vuong's (1989) theories were done to illustrate the steps required to apply their methods. To illustrate these techniques, these methods were applied to the model of Golden *et al.* (1993; also see Golden, in press). This model is misspecified because it is highly constrained by psychologically justifiable constraints determined by modern theories of text knowledge representation. Analysis of the computer simulation results for this particular data set and nonlinear model indicate that (i) classical formulas for the asymptotic covariance matrices provided incorrect estimates (i.e., not internally consistent or consistent with the boot-strap estimates), and (ii) White's (1982, 1989, in press) method provided estimates which were correct (i.e., consistent with the boot-strap estimates).

In conclusion, the methods of making correct statistical inferences in the presence of model misspecification which have been reviewed in this article are important for the following reasons. First, it is likely that the asymptotic approximations of White (1982, 1989, in press) and Vuong (1989) will yield more accurate formulas for deriving between-treatment statistical tests and model selection than classical tests. Second, by not requiring models to satisfy a goodness-of-fit test, the model builder can focus upon the two essential issues of model reliability and model validity in a sequential fashion rather than be forced to consider these issues in parallel as dictated by classical statistical methods. Finally, the general theory of White (1982, 1989, in press) and Vuong (1989) reviewed in this article provides an elegant unified framework for the evaluation and development of new and appropriate statistical tests for nonlinear models of complex phenomena.

APPENDIX 1

The purpose of this section is to provide explicit formulas linking the probability model specified by Eqs. (24)–(28) with White's and Vuong's asymptotic statistical theories.

Notation

Consider the vector-valued quantity $f(x)$ and the scalar-valued quantity $g(x)$ where $f: \mathscr{R}^d \rightarrow \mathscr{R}^d$ and $g: \mathscr{R}^d \rightarrow \mathscr{R}^d$. The quantity df/dx is a matrix whose ijth element is the partial derivative of the ith element of f with respect to the jth element of x. The quantity dg/dx is a vector whose ith element is the partial derivative of the ith element of g with respect to x. The quantity $d^2g/dx^2 = df/dx$ where $f = dg/dx$.

Sample Loss Function

The sample loss function, $\hat{E}_n(\alpha, \lambda)$ for the probability model specified by Eqs. (24)–(28) is given by the formula

$$\hat{E}_n(\alpha, \lambda) = \sum_{i=1}^{n} \hat{E}^i_n(\alpha, \lambda), \tag{32}$$

where

$$\hat{E}^i_n(\alpha, \lambda) = -(1/n) \log[q(x^i \mid x^i(1))]. \tag{33}$$

Substituting (28) into (33),

$$\hat{E}^i_n(\alpha, \lambda) = -(1/n) \log[q(T^i)] \\ -(1/n) \sum_{t=2}^{T^i} \log[x^i(t)^T q^i(t)]. \tag{34}$$

Parameter Estimates

The quasi-maximum likelihood estimate is a critical point of (32). Any critical point of (32) is a parameter vector $(\hat{\alpha}, \hat{\lambda})$ that satisfies (30) and (31). To see this, note that (30) is simply $d\hat{E}_n(\alpha, \lambda)/d\alpha$ set equal to a vector of h zeros, and $\hat{\lambda}$ is obtained by setting the derivative of the error function with respect to λ equal to zero and solving for λ. By the Uniqueness Theorem in Appendix 2, if the Hessian of the error function evaluated at a critical point has only positive eigenvalues, then the obtained critical point is the unique quasi-maximum likelihood estimate.

The first derivative of the sample loss function, $\hat{E}_n(\alpha, \lambda)$, with respect to λ is given by the formula

$$d\hat{E}_n(\alpha, \lambda)/d\lambda = \sum_{i=1}^{n} d\hat{E}^i_n(\alpha, \lambda)/d\lambda,$$

where

$$d\hat{E}^i_n(\alpha, \lambda)/d\lambda = -(1/n)(1 - T^i/\lambda). \tag{35}$$

Now (31) is obtained by substituting $\hat{\lambda}$ into the equation

$$d\hat{E}_n(\alpha, \lambda)/d\lambda = 0$$

and solving for $\hat{\lambda}$. The first derivative of the sample loss function, $\hat{E}_n(\alpha, \lambda)$, with respect to the h-dimensional α vector is given by the formula

$$d\hat{E}_n(\alpha, \lambda)/d\alpha = \sum_{i=1}^{n} d\hat{E}^i_n(\alpha, \lambda)/d\alpha,$$

where

$$d\hat{E}^i_n(\alpha, \lambda)/d\alpha = -(1/n) \sum_{t=2}^{T^i} (x^i(t) - q^i(t))^T v^i(t). \tag{36}$$

Asymptotic Covariance Matrices of the Parameter Estimates

Let $(\hat{\alpha}, \hat{\lambda})$ be a critical point of the sample loss function defined according to (30) and (31). Let $\hat{z}^i$ be the column vector z^i defined by

$$z^i = [d\hat{E}^i_n(\alpha, \lambda)/d\alpha \quad d\hat{E}^i_n(\alpha, \lambda)/d\lambda]^T \tag{37}$$

and evaluated at $(\hat{\alpha}, \hat{\lambda})$. Let the matrix $\hat{Z}_n = [\hat{z}^1 \, \hat{z}^2 \cdots \hat{z}^n]$.

Then matrix $\hat{B}_n$ as in (9) is then

$$\hat{B}_n = n\hat{Z}^T_n \hat{Z}_n.$$

Moreover, let $\hat{E}^f_n$ be the sample loss function associated with probability model $F_W = \{q_f(\cdot\,; w)\}$, and $\hat{E}^g_n$ be the sample loss function associated with probability model $G_Y = \{q_g(\cdot\,; y)\}$. Let the matrix $\hat{Z}_n$ associated with model F_W be indicated by $\hat{Z}_f$ and let the matrix $\hat{Z}_n$ associated with model G_Y be indicated by $\hat{Z}_g$. Then

$$\hat{B}^{fg}_n = n\hat{Z}^T_g \hat{Z}_f.$$

Let D_{qi} be a diagonal matrix whose jth on-diagonal element is the jth element of vector $q^i(t)$. Let 0_h be a h-dimensional column vector of zeros. Then note that $d^2\hat{E}_n(\alpha, \lambda)/d\lambda \, d\alpha = 0_h$ for any $(\alpha, \lambda) \in \mathscr{R}^{h+1}$. The Hessian of $\hat{E}_n(\alpha, \lambda)$, $\nabla^2 \hat{E}_n(\alpha, \lambda)$ is then given by the formula

$$\nabla^2 \hat{E}_n(\alpha, \lambda) = \sum_{i=1}^{n} \begin{bmatrix} d^2 E^i_n(\alpha, \lambda)/d\alpha^2 & 0_h \\ 0^T_h & d^2 E^i_n(\alpha, \lambda)/d\lambda^2 \end{bmatrix}. \tag{38}$$

The quantity $d^2 E^i_n(\alpha, \lambda)/d\alpha^2$ is now computed. Taking the derivative of (36) with respect to α,

$$d^2 E^i_n(\alpha, \lambda)/d\alpha^2 = -(1/n) \sum_{t=2}^{T^i} (0 - [dq^i(t)/d\alpha]^T v^i(t)). \tag{39}$$

Then note that

$$\begin{aligned} dq^i(t)^T/d\alpha &= [dq^i(t)/dh^i(t)][dh^i(t)/d\alpha] \\ &= [D_{qi} - q^i(t)\, q^i(t)^T]\, v^i(t). \end{aligned} \tag{40}$$

Thus,

$$\begin{aligned} &d^2 E^i_n(\alpha, \lambda)/d\alpha^2 \\ &\quad = (1/n) \sum_{t=2}^{T^i} [v^i(t)]^T [D_{qi} - q^i(t)\, q^i(t)^T]\, v^i(t) \end{aligned}$$

and

$$d^2 E^i_n(\alpha, \lambda)/d\lambda^2 = T^i/(n\lambda^2).$$

The matrix $-\hat{A}_n$ is computed as in (8) by evaluating -1 times (38) at the critical point $(\hat{\alpha}, \hat{\lambda})$ where $(\hat{\alpha}, \hat{\lambda})$ is computed from (30) and (31). White's asymptotic covariance matrix $\hat{C}_n$ may then be computed using (7) from $\hat{A}_n$ and $\hat{B}_n$.

APPENDIX 2

The purpose of this appendix is to state and prove explicit conditions for the parameter estimates of the proposed PDP model to be unique. To prove this theorem, it is first necessary to prove two lemmas.

LEMMA 1. *Let A be a real-valued symmetric positive semi-definite d-dimensional matrix. Let X be a matrix, then the matrix $X^T AX$ is positive semi-definite.*

Proof. Since A is a d-dimensional real symmetric positive semi-definite matrix,

$$A = \sum_{i=1}^{d} \lambda_i e_i e^T_i,$$

where e_i is the real-valued column eigenvector associated with the ith non-negative eigenvalue λ_i of A. The matrix $X^T AX$ is positive semi-definite if and only if for any real-valued column vector y, $y^T(X^T AX)\, y \geqslant 0$. But

$$\begin{aligned} y^T(X^T AX)\, y &= \sum_{i=1}^{d} \lambda_i (e^T_i Xy)(e^T_i Xy) \\ &= \sum_{i=1}^{d} \lambda_i (e^T_i Xy)^2 \geqslant 0. \qquad \text{Q.E.D.} \end{aligned}$$

LEMMA 2. *Let M be a d-dimensional diagonal square matrix whose ith on-diagonal element, m_{ii} is given by*

$$m_{ii} = p_i - p^2_i$$

and whose ijth off-diagonal element, m_{ij}, is given by

$$m_{ij} = -p_i p_j,$$

where $0 < p_i < 1$ and $\sum^d_{i=1} p_i = 1$. Then M is positive semi-definite.

Proof. By Gershgorin's circle theorem (Noble and Daniel, 1977, p. 289) every eigenvalue λ of M must satisfy at least one of the inequalities:

$$|\lambda - m_{ii}| \leqslant \sum_{j \neq i} |m_{ij}|. \tag{41}$$

Introducing the assumed constraints on m_{ii} and m_{ij}, (41) becomes

$$|\lambda - p_i + p_i^2| \leqslant \sum_{j \neq i} p_i p_j = p_i \sum_{j \neq i} p_j = p_i(1 - p_i).$$

Thus, (41) subject to the constraints on M becomes

$$-p_i(1-p_i) + p_i - p_i^2 \leqslant \lambda \leqslant p_i(1-p_i) + p_i - p_i^2,$$

which simplifies to

$$0 \leqslant \lambda \leqslant 2p_i(1-p_i).$$

Thus, all eigenvalues of M are greater than or equal to zero. Q.E.D.

THEOREM 3. *Let $\hat{E}_n(\alpha, \lambda)$ be the Hessian of the sample loss function defined by* (38) *in Appendix* 1. *Let $(\hat{\alpha}, \hat{\lambda})$ be a critical point of $\hat{E}_n(\alpha, \lambda)$ which is computed using* (30) *and* (31). *If the eigenvalues of the Hessian of $\hat{E}_n(\alpha, \lambda)$ evaluated at $(\hat{\alpha}, \hat{\lambda})$ are strictly positive, then $(\hat{\alpha}, \hat{\lambda})$ is the unique quasi-maximum likelihood estimate* (*i.e., strict global minimum of $\hat{E}_n(\alpha, \lambda)$*).

Proof. First note that

$$d^2 \hat{E}^i(\alpha, \lambda)/d\lambda^2 = T^i/(N\lambda^2) \geqslant 0$$

for any $(\alpha, \lambda) \in \mathscr{R}^{h+1}$. Thus, $d^2 \hat{E}_n(\alpha, \lambda)/d\lambda^2$ is non-negative.

Second, it now is shown that $d^2 E^N(\alpha, \lambda)/d\alpha^2$ is positive semi-definite for any $(\alpha, \lambda) \in \mathscr{R}^{h+1}$. By Proposition 2, the matrix $D_{qi} - q^i(t)\, q^i(t)$ is positive semi-definite. And this observation in conjuction with Proposition 1 shows that

$$d^2 \hat{E}^i(\alpha, \lambda)/d\alpha^2$$
$$= (1/N) \sum_{t=2}^{T^i} [v^i(t)]^T [D_{qi} - q^i(t)\, q^i(t)]\, v^i(t)$$

is positive semi-definite for any $(\alpha, \lambda) \in \mathscr{R}^{h+1}$. Thus, $d^2 \hat{E}^n(\alpha, \lambda)/d\alpha^2$ is positive semi-definite for any $(\alpha, \lambda) \in \mathscr{R}^{h+1}$ since the sum of a finite number of positive semi-definite matrices is a positive semi-definite matrix.

Inspection of $\nabla^2 \hat{E}_n(\alpha, \lambda)$ in (38) Appendix 1 in conjunction with the above two results show that for any $(\alpha, \lambda) \in \mathscr{R}^{h+1}$, $\nabla^2 \hat{E}_n(\alpha, \lambda)$ in (38) is positive semi-definite for any $(\alpha, \lambda) \in \mathscr{R}^{h+1}$. Since $\hat{E}_n(\alpha, \lambda)$ has continuous second partial derivatives and is positive semi-definite for all $(\alpha, \lambda) \in \mathscr{R}^{h+1}$, then $\hat{E}_n(\alpha, \lambda)$ is convex on $\mathscr{R}^{h+1}$ (Luenberger, 1984, p. 180). Moreover, since $\hat{E}_n(\alpha, \lambda)$ is convex on $\mathscr{R}^{h+1}$, any strict local minimum of $\hat{E}_n(\alpha, \lambda)$ is the unique strict global minimum (Luenberger, 1984, p. 181). Q.E.D.

APPENDIX 3

The purpose of this appendix is to review the basic steps of the variance test described by Vuong (1989). Let $\hat{A}_n^f$ be the Hessian matrix associated with model F_W as in (8). Let $\hat{B}_n^f$ be the matrix associated with model F_W as in (9). Let $\hat{A}_n^g$ be the Hessian matrix associated with model G_Y as in (8). Let $\hat{B}_n^g$ be the matrix associated with model G_Y as in (9). Also define a new matrix, $\hat{B}_n^{fg}$, so that

$$\hat{B}_n^{fg} = n^{-1} \sum_{i=1}^{n} \nabla_w \log[q_f(x_i; \hat{w}_n)]\, \nabla_y \log[q_g(x_i; \hat{y}^n)]^T, \tag{42}$$

where the notation $\nabla_y \log[q_g(x_i; \hat{y}_n)]$ indicates that the gradient of $\log[q_g(x_i; y)]$ with respect to y should be computed and then evaluated at the point $\hat{y}_n$. Then $\hat{R}_n$ is defined as follows

$$\hat{R}_n = \begin{bmatrix} -\hat{B}_n^f(\hat{A}_n^f)^{-1} & -\hat{B}_n^{fg}(\hat{A}_n^g)^{-1} \\ (\hat{B}_n^{fg})^T(\hat{A}_n^f)^{-1} & \hat{B}_n^g(\hat{A}_n^g)^{-1} \end{bmatrix} \tag{43}$$

The first stage of Vuong's (1989) model selection procedure may now be reviewed. As previously noted, Vuong (1989) makes a number of regularity assumptions which are similar to the regularity assumptions associated with White's asymptotic statistical theory (see Vuong, 1989, for additional details). Given these assumptions, Vuong (1989) shows that as n increases, $\hat{\sigma}_{vn}^2 \to \sigma_v^2$ with probability one. Vuong (1989, Lemma 4.1) also showed that the null hypothesis that F_W and G_Y are *not* strictly non-nested (i.e., $H_0: q_f(\cdot; w^*) = q_g(\cdot; y^*)$) is formally equivalent to the null hypothesis that $\sigma_v^2 = 0$. Vuong (1989) then showed that the variance statistic $\hat{\sigma}_{vn}^2$ has a weighted chi-square distribution (see Appendix 4 for details regarding this distribution; also Vuong, 1989). Thus, Vuong (1989) was able to construct a statistical test based upon the variance statistic, $\hat{\sigma}_{vn}^2$, in order to decide if two probability models F_W and G_Y are strictly non-nested.

This variance test consists of the following three step procedure. First, compute quasi-maximum likelihood estimates $\hat{y}_n$ and $\hat{w}_n$ for each of the two probability models G_Y and F_W. Second, let $\hat{\omega}^2$ be a vector whose ith element is the square of the ith eigenvalue of $\hat{R}_n$. Using the algorithm in Appendix 4, compute the critical value $\sigma_\alpha^2(\hat{\omega}^2)$ which has the property that a weighted chi-square random variable with parameter vector $\hat{\omega}^2$ will exceed $\sigma_\alpha^2(\hat{\omega}^2)$. And third, decide the models G_Y and F_W are strictly non-nested (i.e., reject $H_0 : D = 0$) if $n\hat{\sigma}_{vn}^2 > \sigma_\alpha^2(\hat{\omega}^2)$.

APPENDIX 4

The purpose of this Appendix is to define a cumulative weighted chi-square distribution following Vuong (1989) and suggest a numerically efficient way to calculate this distribution. The proposed method for calculating this distribution was suggested by Stuart Golden (personal communication, December 1992).

Definition 4. Let $Z = [Z_1, ..., Z_m]$ be a vector of m independent and identically distributed Gaussian random variables with mean zero and variance one. Then the random variable $U_m = \sum_{i=1}^{m} k_i Z_i^2$ has a weighted chi-square distribution with weighting parameter vector $K = [k_1, ..., k_m]$.

The probability density function of $y_i = Z_i^2$ where Z_i is a Gaussian random variable with mean zero and variance one is called a chi-square density function with one degree of freedom. Let j be the square root of negative one. The characteristic function of y_i is defined as the expectation of $\exp(jty_i)$ with respect to the probability distribution of y_i. Wilks (1962, p. 183) notes that the characteristic function of the chi-square random variable $y_i = Z_i^2$ is given by

$$\Phi_{y_i}(t) = [1 - 2jt]^{-1/2}. \tag{44}$$

The characteristic function of

$$U_m = \sum_{i=1}^{m} k_i y_i$$

is now computed using a theorem discussed by Wilks (1962, p. 121). In particular,

$$\Phi_{U_m}(t) = \prod_{i=1}^{m} \Phi_{y_i}(k_i t), \tag{45}$$

where $\Phi_{y_i}(t)$ is the characteristic function of y_i since $y_1, ..., y_m$ are independent random varibles.

Given the characteristic function of U_m, it is now desired to compute the cumulative distribution function, $F(U_m)$, of U_m. Making use of the fact that U_m is always non-negative and the "inversion formula" provided by Manoukian (1986, p. 12), we have

$$F(U_m) = (1/2\pi) \int_{-\infty}^{\infty} [1 - \exp(-jU_m t)](\Phi_{U_m}(t)/jt)\, dt. \tag{46}$$

The integral in (46) is then numerically evaluated by making the substitution of variables

$$t_n = (2\pi/T)(n/N),$$

where T is the "sampling rate period" and N is the number of sampling points. The integral in (46) then becomes

$$F(U_m) = (1/TN) \int_{-\infty}^{\infty} [1 - \exp(-jU_m t_n)](\Phi_{U_m}(t_n)/jt_n)\, dn. \tag{47}$$

The motivation for this substitution of variables is based upon the idea that the integrand of (46) is modelled as a complex periodic function whose period just happens to be equal to NT.

The integral in (47) is then approximated as a sum to obtain

$$F(U_m) = (1/NT) \sum_{n=-(N/2)+1}^{N/2} [1 - \exp(-jU_m t_n)](\Phi_{U_m}(t_n)/jt_n), \tag{48}$$

where $t_n = 2\pi/(TN)$.

By choosing K to be a vector of d ones, $F(U_m)$ reduces to the familiar chi-square cumulative distribution function associated with a chi-square random variable with d degrees of freedom. Using $N = 50{,}0000$ and $T = 0.0001$, an informal comparison between this proposed numerical algorithm and published chi-square tables showed a relative error of approximately 1%. If desired, the accuracy of this algorithm could be improved by evaluating (46) using more sophisticated numerical integration methods. The essential trick is the expression of the cumulative distribution function $F(U_m)$ as the single integral in (46).

REFERENCES

Allison, P. D., & Liker, J. K. (1982). Analyzing sequential categorical data on dyadic interaction: A comment on Gottman. *Psychological Bulletin*, **91**, 393–403.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, **31**, 419–456.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Falmagne, J., Koppen, M., Villano, M., Doigonon, J., & Johannsen, L. (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review*, **97**, 201–224.

Golden, R. M. (1988a). Probabilistic characterization of neutral model computations. In D. Z. Anderson (Ed.), *Neural networks and information processing* (pp. 310–316). New York: American Institute of Physics.

Golden, R. M. (1988b). Relating neural networks to traditional engineering approaches. In *The proceedings of the artificial intelligence and advanced computer technology conference*. Glen Ellyn, IL: Tower Conference Management Company.

Golden, R. M. (1988c). A unified framework for connectionist systems. *Biological Cybernetics*, **59**, 109–120.

Golden, R. M. (in press). Analysis of categorical time-series text recall data using a connectionist model. *Journal of Biological Systems*.

Golden, R. M., Golden, S. F., Strickland, J., & Choi, I. (1983). A psychometric pdp model of temporal structure in story recall. In *The proceedings of the fourteenth annual conference of the cognitive science society* (pp. 487–491). Hillsdale, NJ: Erlbaum.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium in mathematical statistics and probability*. Berkeley: University of California Press.

Jordan, M. I. (1992). Constrained supervised learning. *Journal of Mathematical Psychology*, **36**, 396–425.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.

Levinson, S. E., Ljolje, A., & Miller, L. G. (1988). Large vocabulary speech recognition using a hidden markov model for acoustic/phonetic classification. In 1988 *IEEE International conference on acoustics, speech, and signal processing* (Vol. 1, pp. 505–508). Piscataway, NJ: IEEE Service Center.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Luenberger, D. G. (1984). *Linear and nonlinear programming*. Reading, MA: Addison–Wesley.

Manoukian, E. B. (1986). *Modern concepts and theorems of mathematical statistics*. New York: Springer-Verlag.

Montgomery, D. C., & Peck, E. A. (1982). *Introduction to linear regression analysis*. New York: Wiley.

Noble, B., & Daniel, J. W. (1977). *Applied linear algebra*. Englewood Cliffs, NJ: Prentice–Hall.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, **95**, 318–339.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1: Foundations, pp. 318–362). Cambridge, MA: MIT Press.

Trees, H. L. V. (1968). *Detection, estimation, and modulation theory*. New York: Wiley.

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. (1989). Learning in artificial neutral networks: A statistical perspective. *Neural Computation*, **1**, 425–464.

White, H. (in press). *Estimation, inference, and specification analysis*. New York: Cambridge Univ. Press.

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

Wilks, S. S. (1962). *Mathematical statistics*. New York: Wiley.

Received: January 31, 1992