Golden, R. M. (2006). Technical Report: Knowledge Digraph Contribution Analysis (Version 01-29-06). School of Behavioral and Brain Sciences (GR4.1), University of Texas at Dallas, Richardson, TX 75083-0688.

Technical Report: Knowledge Digraph Contribution Analysis

Richard M. Golden (golden@utdallas.edu)

Classical sequential data analysis methods are not typically used for data analysis purposes since they tend to be better suited for exploratory rather than confirmatory data analysis. In particular, such methods do not incorporate or explicitly encourage theoretician-imposted constraints upon patterns of assocative strengths. Golden (1995, 1998) has developed a highly constrained parametric multinomial time-series regression model for categorical time-series analysis of free response data as an ordered sequence of propositions. Golden (1998) refers to models of this type as Knowledge Digraph Contribution (KDC) analysis models since the researcher specifies a collection of directed graphs representing different types of semantic relations among propositions and then the KDC program estimates a contribution weight parameter for each directed graph (digraph). The directed graphs are based upon theories of semantic connectivity. Additionally, KDC analysis has a distinct advantage over classical sequential data analysis methods because all of the asymptotic statistical tests developed using KDC analysis are derived within the general theory of model misspecification which permits reliable statistical inferences even when the theoretical assumptions about the types of semantic relations among propositions are not entirely correct (see White, 1982, 1994; Golden, 1995, 1996, 2003; for relevant reviews).

The purpose of this technical report is to specify the key mathematical assumptions of KDC (Knowledge Digraph Contribution) analysis.

1 Mathematical Theory

1.1 Knowledge Digraphs

Data in knowledge digraph analysis consists of observing one or more sequences of "items". For example, an "item" may be a phoneme, a participant response, a visual image, or an action. Assume there are a finite number of d items which may be observed in a particular sequence of observations. It will mathematically convenient to refer to the ith item as $\mathbf{s}_i$ where $\mathbf{s}_i$ is the ith column of a d-dimensional identity matrix, $i = 1, \ldots, d$.

Let the set of all possible items $\Omega \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_d\}$. Let $\mathbf{0}_{m\times n}$ be a matrix of zeros with m rows and n columns. Let $\mathbf{0}_d \equiv \mathbf{0}_{d\times 1}$. It will also be mathematically convenient to define: $\Gamma \equiv \Omega \cup \{\mathbf{0}_d\}$.

Let $v \in \{1, \ldots, q\}$. Denote a *temporal knowledge digraph* of *knowledge type* v as a k-tuplet: $(\mathbf{D}^v[1], \ldots, \mathbf{D}^v[k])$ where $\mathbf{D}^v[m] \in \mathcal{R}^{d\times d}$ is called a *knowledge digraph* of *knowledge type* v with *lag* m for $m = 1, \ldots, k$. The element in row i and column j of $\mathbf{D}^v[m]$, $D^v[m]ij$, is called the ijth *digraph link weight* for the knowledge digraph $\mathbf{D}^v[m]$ with lag m. A larger value of the digraph link weight $D^v[m]_{ij}$ corresponds to the hypothesis that: Subsequences where item $\mathbf{s}_j$ is followed subsequently by $m-1$ items, and then immediately by item $\mathbf{s}_i$ are likely to occur in the observed data.

For example, let $\mathbf{D}^v[m]$ be a knowledge digraph of type v with lag m which is defined such that:

$$\mathbf{D}^v[m] = \mathbf{s}_2(\mathbf{s}_1)^T + \mathbf{s}_2(\mathbf{s}_3)^T + \mathbf{s}_5\mathbf{s}_2^T + \mathbf{s}_6\mathbf{u}\mathbf{s}_2^T + \mathbf{s}_4\mathbf{s}_5^T + \mathbf{s}_4\mathbf{s}_6^T.$$

The digraph $\mathbf{D}^v[m]$ is depicted graphically in Figure 1. Referring to Figure 1, suppose that $m = 1$. A digraph $\mathbf{D}^v[1]$ of type v and lag 1 represents a hypothesis regarding the presence of sequences in the observed data such as: $\{1, 2, 6, 4\}$, $\{3, 2, 5, 4\}$, and $\{3, 2, 6, 4\}$. Now suppose that the digraph in Figure 1 was a lag 2 digraph so that $m = 2$. Then, $\mathbf{D}^v[2]$ would correspond to a hypothesis regarding the presence of sequences in the observed data such as: $\{1, x_4, 2, x_5, 6, x_6, 4\}$, $\{3, x_7, 2, x_8, 5, x_9, 4\}$, and $\{3, x_{10}, 2, x_{11}, 6, x_{12}, 4\}$ where $x_i \in \Omega$ for $i = 1, \ldots, 12$.

Define $\mathcal{D}^v \equiv (\mathbf{D}^v[1], \ldots, \mathbf{D}^v[k]) \in \mathcal{R}^{d\times dk}$ for $v = 1, \ldots, q$. The set $\{\mathcal{D}^1, \ldots, \mathcal{D}^q\}$ is called a *knowledge digraph model*. The *knowledge digraph connection function* $\mathbf{W} : \mathcal{R}^q \to \mathcal{R}^{d\times dk}$ such that:

$$\mathbf{W}(\theta) \equiv \sum_{v=1}^{q} \theta_v \mathcal{D}^v$$

where $\theta \equiv [\theta_1, \ldots, \theta_q] \in \Theta \subseteq \mathcal{R}^q$.

1.2 Data Generating Processes

A stochastic process $\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, \ldots$ is called *τ-dependent* if $\tilde{\mathbf{f}}_i$ and $\tilde{\mathbf{f}}_j$ are independent for all $|i-j| > \tau$. Thus, if $\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, \ldots$ is a τ-dependent stochastic process where $\tau = 0$, then $\tilde{\mathbf{f}}_i$ and $\tilde{\mathbf{f}}_j$ are statistically independent for all $i \neq j$.

A stochastic process $\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, \ldots$ is called *strictly stationary* if the joint distribution of $\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, \ldots$ is identical to the joint distribution of $\tilde{\mathbf{f}}_{m+1}, \tilde{\mathbf{f}}_{m+2}, \ldots$ for $m = 1, 2, \ldots$. Observations in strictly stationary stochastic processes are always identically distributed. Note that a strictly stationary stochastic process which is *τ-dependent* where $\tau = 0$ is a stochastic process consisting of independent and identically distributed observations.

A d-dimensional random vector $\tilde{\mathbf{f}}$ will be called *categorical* if the range of $\tilde{\mathbf{f}}$ is finite.

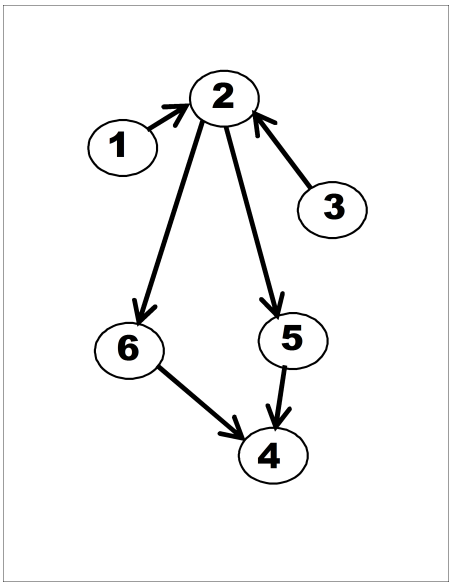

Fig. 1. *Figure 1.* A lag k knowledge digraph of type v which is denoted by $\mathbf{D}^v[k]$ is graphically represented as six nodes with an arrow connecting node j to node i with a digraph link weight equal to the ijth element of $\mathbf{D}^v[k]$. By convention, no arrow is drawn for a digraph link weight if that digraph link weight is equal to zero.

Assumption A1. Data Generating Process Specification. Let τ be a finite non-negative integer. The observed data are a realization of a strictly stationary τ-dependent stochastic process of d-dimensional categorical random vectors: $\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, \ldots$.

Assumption A1 is a relatively weak assumption regarding the nature of the stochastic process which generates the observed categorical time-series data.

In a typical application of KDC analysis, $\tilde{\mathbf{f}}_t^k$ models the tth response from participant k in a group of n participants, $k = 1, \ldots, T_k$, $n = 1, \ldots, k$. In this situation, it is assumed that between-participant responses are statistically independent (i.e., $\tilde{\mathbf{f}}_t^k$ and $\tilde{\mathbf{f}}_s^j$ are statistically independent for all $k \neq j$ and for all t, s). Observations within a sequence are assumed to be statistically independent if they are sufficiently separated in time as specified by the τ assumption (i.e., $\tilde{\mathbf{f}}_t^k$ and $\tilde{\mathbf{f}}_s^k$ are statistically independent if $|k - s| > \tau$).

Finally, let $h : \mathcal{R}^d \rightarrow \mathcal{R}$. Throughout this article, the notation $E\{h(\tilde{\mathbf{f}})\}$ refers to the expected value of $h(\tilde{\mathbf{f}})$ with respect to the distribution of $\tilde{\mathbf{f}}$ (when it exists). For strictly stationary stochastic process, $E\{h(\tilde{\mathbf{f}}_t)\} = E\{h(\tilde{\mathbf{f}}_m)\}$ for all $t, m = 1, 2, \ldots$.

1.3 KDC Probability Model

A KDC probability model is a set of specifications which (ideally) contains the specification for the probability mass function (pmf) which generated the observed data according to Assumption A1. Following White (1982), we make a clear distinction between the assumptions associated with the data

generating process (DGP) and the assumptions associated with the researcher's model of the DGP. If the KDC probability model contains the pmf specification which generated the observed data, then the KDC probability model is said to be *correctly specified* with respect to the DGP. If the KDC probability model does not contain the pmf specification which generated the observed data, then the KDC probability model is said to be *misspecified* with respect to the DGP. The asymptotic theory of statistical inference developed here supports reliable statistical inferences in the presence of model misspecification (see White, 1982; Golden, 1995, 1996, for relevant reviews).

Let $\mathbf{u}_t \equiv [\mathbf{f}_{t-1}, \ldots, \mathbf{f}_{t-\tau}]$. The function $\mathbf{exp} : \mathcal{R}^d \rightarrow (0, \infty)^d$ is defined such that the ith element of the d-dimensional column vector $\mathbf{exp}([x_1, \ldots, x_d]^T)$ is equal to $exp(x_i)$, $i = 1, \ldots, d$. For $t = 1, 2, \ldots$, let $\mathbf{h}_t(\mathbf{u}_t; \theta) \equiv \mathbf{W}_\theta \mathbf{u}_t$ and

$$\mathbf{p}_t(\mathbf{u}_t; \theta) \equiv \frac{\mathbf{exp}(\mathbf{h}_t(\mathbf{u}_t; \theta))}{\mathbf{1}_d^T \mathbf{exp}(\mathbf{h}_t(\mathbf{u}_t; \theta))}. \tag{1}$$

Assumption A2. Knowledge Digraph Probability Model Specifications. Let Θ be a compact and non-empty subset of a q-dimensional real vector space. Let $\mathbf{p}_t$ be defined as in (1). Let $p : \Omega \times \Gamma^\tau \times \Theta \rightarrow [0, 1]$ be defined for all $\mathbf{f}_t \in \Omega$ and for all $\mathbf{u}_t \in \Gamma^\tau$ such that for $t = 1, 2, \ldots$:

$$p(\mathbf{f}_t | \mathbf{u}_t; \theta) = \mathbf{f}_t^T \mathbf{p}_t(\mathbf{u}_t; \theta)$$

for all $\theta \in \Theta$.

A set $\mathcal{M} \equiv \{\mathbf{p}_t : \mathbf{p}_t(\cdot; \theta), \theta \in \Theta\}$ whose elements satisfy A2 is called a *knowledge digraph probability model*. Also note that Assumption A2 may be interpreted as specifying a type of multinomial logistic time-series model. Assumption A2 may also be interpreted as specifying a type of connectionist time-series model as shown in Figure 2.

Assumption A3. Multivariate Gaussian Prior Specifications. Let $p_\theta : \Theta \times \Theta \times \Theta^2 \rightarrow [0, \infty)$ be defined such that $p_\theta(\cdot; \mathbf{m}_\theta, \mathbf{C}_\theta)$ is a multivariate Gaussian density with mean $\mathbf{m}_\theta$ and covariance matrix $\mathbf{C}_\theta$ for all $\mathbf{m}_\theta \in \Theta$ and for all positive definite $\mathbf{C}_\theta \in \Theta^2$.

1.4 Parameter Estimation

Parameter estimation requires the specification of an objective function whose global minima may be identified as the parameter estimates. The objective function used in KDC analysis is motivated by noting that:

$$p(\theta | \mathbf{f}_1, \ldots, \mathbf{f}_n) \equiv \left[\prod_{t=1}^{n} p\left(\mathbf{f}_t | \mathbf{u}_t; \theta\right) \right] p_\theta(\theta). \tag{2}$$

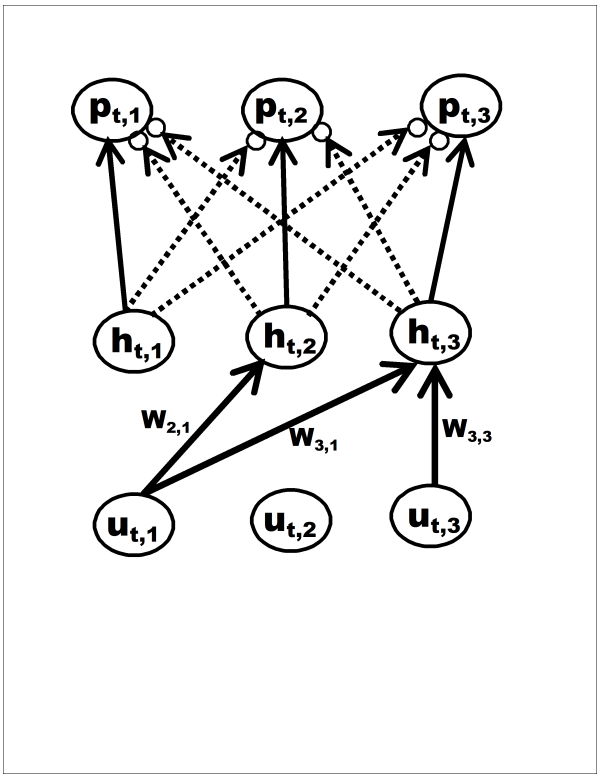

Fig. 2. *Figure 2.* Connectionist network interpretation of KDC probability model. Note that the activation pattern over the *input units* correspond to $\mathbf{u}_t \equiv [u_{t,1}u_{t,2}u_{t,3}]$, the activation pattern over the *hidden units* correspond to $\mathbf{h}_t \equiv [h_{t,1}h_{t,2}h_{t,3}]$, and the activation pattern over the *output units* is $\mathbf{p}_t \equiv [p_{t,1}p_{t,2}p_{t,3}]$. The *connection weights* from the input units to the hidden units are specified by $\mathbf{W}(\theta)$. The non-modifiable nonlinear transformation mapping $\mathbf{h}_t$ into $\mathbf{p}_t$ is defined by Equation (1).

The global maxima of $p(\theta|\mathbf{f}_1, \ldots, \mathbf{f}_n)$ thus may be identified as corresponding to MAP (Maximum A Posteriori) estimates.

Let

$$Z_\theta \equiv (2\pi)^{q/2}(det\mathbf{C}_\theta)^{1/2}.$$

Let the *loss function* $c^\Theta : \Theta \times \mathcal{R}^{d(\tau+1)} \rightarrow \mathcal{R}$ be defined such that:

$$c(\theta; \mathbf{f}_t, \mathbf{f}_{t-1}, \ldots, \mathbf{f}_{t-\tau}) = -\log\left(p(\theta|\mathbf{f}_1, \ldots, \mathbf{f}_n)\right) + logZ_\theta$$

$$= -\log\left(\mathbf{f}_t^T \log(\mathbf{p}_t(\mathbf{u}_t; \theta)\right) + (1/2)(\theta - \mathbf{m}_\theta)^T[\mathbf{C}_\theta]^{-1}(\theta - \mathbf{m}_\theta)$$

Let

$$l_n^\Theta(\theta; \mathbf{f}_1, \ldots, \mathbf{f}_n) = n^{-1}\sum_{i=1}^{n} c^\Theta(\theta; \mathbf{f}_1, \ldots, \mathbf{f}_n) - n^{-1}log(p_\theta(\theta)) \tag{3}$$

where the q-dimensional real vector $\mathbf{m}_\theta$ and the q-dimensional positive definite symmetric matrix $\mathbf{C}_\theta$ are known constants.

First note, that the global minima of l_n^{Θ} correspond to the global maxima of $p(\theta|\mathbf{f}_1,\ldots,\mathbf{f}_n)$. In addition, the global minima of the first term of (3) (which is typically referred to as tne negative log-likelihood function) correspond to *maximum likelihood estimates*. Note that as n becomes large, the second term in (3) becomes small relative to the log-likelihood term implying that under general conditions MAP estimation and ML estimation will be asymptotically equivalent. In addition to having important semantic properties, the function l_n has important computational properties as well which are summarized in the following Theorem T1.

Theorem T1: KDC MAP Objective Function Properties. The objective function l_n in (3) is convex and analytic on $\mathcal{R}^q$. In addition, if $\nabla^2 l_n$ evaluated at a critical point, $\hat{\theta}_n$, has strictly positive eigenvalues, then $\hat{\theta}_n$ is the unique global minimum of l_n.

Theorem T1 establishes that l_n is a convex differentiable function standard nonlinear optimization methods such as the Newton-Raphson algorithm with an appropriate linesearch may be designed to find a global minimum, $\hat{\theta}_n$, of l_n. Finally, Theorem T1 provides a simple computational test for determining if $\hat{\theta}_n$ is the unique global minimum of l_n.

The function l_n^{Θ} may be viewed as a realization of the "random function"

$$\tilde{l}_n^{\Theta}(\cdot) \equiv l_n(\cdot;\tilde{\mathbf{f}}_1,\ldots,\tilde{\mathbf{f}}_n).$$

Let the *MAP estimate* $\tilde{\theta}_n$ be defined (when it exists) as a global minimum of $\tilde{l}_n^{\Theta}$. Suppose the expected value of $\tilde{l}_n^{\Theta}$, $l^{\Theta}:\theta\rightarrow\mathcal{R}$, exists and there is a unique global minimum of l^{Θ}. This unique global minimum of l is called the *pseudo-true parameter* since it can be shown (Kullback-Leibler, 1959) that this global minimum corresponds to the true model parameters if the probability model is correctly specified. If the probability model is not correctly specified, then the global minimum of l may be interpreted as minimizing the cross-entropy between the KDC probability model and the DGP.

Some notation is now introduced for the purposes of stating some theorems which explicitly relate properties of $\tilde{l}^{\Theta}$ to properties of l_n^{Θ}. Let the *gradient per observation* $\mathbf{g}_i^{\Theta} \equiv -\log\left(p(\mathbf{f}_i|\mathbf{f}_{i-1},\ldots,\mathbf{f}_{i-\tau};\theta)\right)$. Let the *sample average gradient* $\mathbf{g}^{\Theta} \equiv n^{-1}\sum_{i=1}^{n}\mathbf{g}_i^{\Theta}$. Let $J_{i,\tau} \equiv \{j\in\{1,2,\ldots\} : |i-j|\leq\tau\}$. Let the *sample Outer Product Gradient (OPG) Hessian Estimator*

$$\mathbf{B}_n^{\Theta} \equiv n^{-1}\sum_{i=1}^{n}\sum_{j\in J_{i,\tau}}\mathbf{g}_i^{\Theta}(\mathbf{g}_j^{\Theta})^T.$$

The *sample Sandwich Estimator* $\mathbf{C}_n^{\Theta} \equiv [\mathbf{A}_n^{\Theta}]^{-1}\mathbf{B}_n^{\Theta}[\mathbf{A}_n^{\Theta}]^{-1}$ when it exists. Let $\mathbf{A}^{\Theta} \equiv E\{\tilde{\mathbf{A}}_n^{\Theta}\}$, $\mathbf{B}^{\Theta} \equiv E\{\tilde{\mathbf{B}}_n^{\Theta}\}$, $\mathbf{C}^{\Theta} \equiv E\{\tilde{\mathbf{C}}_n^{\Theta}\}$, and $\mathbf{l}^{\Theta} \equiv E\{\tilde{\mathbf{l}}_n^{\Theta}\}$. Let $\mathbf{g}^{\Theta*}$, $\mathbf{A}^{\Theta*}$, $\mathbf{B}^{\Theta*}$, and $\mathbf{C}^{\Theta*}$ be defined respectively as $\mathbf{g}^{\Theta}$, $\mathbf{A}^{\Theta}$, $\mathbf{B}^{\Theta}$, and $\mathbf{C}^{\Theta}$ evaluated at θ^*.

Assumption A4. Unique Local Minimum. Assume $\mathbf{A}^{\Theta}(\theta^*)$ is positive definite.

Assumption A4 guarantees that θ^* is a strict local minimum. Since Theorem T1 shows that l^{Θ} is convex for KDC probability models, A4 and T1 together imply that any strict local minimum θ^* which satisfies

A4 must be the unique global minimum. Although this establishes global identifiability of the KDC probability models, the following Theorem T2 provides a stronger result.

Theorem T2: Consistent Estimation of Parameter Estimates. Assume A1, A2, A3, and A4 hold. Then $\tilde{\theta}_n$ exists for $n = 1, 2, \ldots$. In addition, as $n \to \infty$, $\tilde{\theta}_n \to \theta^*$ w.p.1.

Theorem T2 thus establishes the existence, uniqueness, and consistency of the parameter estimates for a KDC probability model provided that Assumption A4 is satisfied.

The asymptotic distribution of $\tilde{\theta}_n$ is now characterized.

Assumption A5. OPG Hessian Positive Definiteness. Assume $\mathbf{B}^{\Theta *}$ is positive definite.

Theorem T3: Asymptotic Distribution of Estimates. Assume A1, A2, A3, A4, and A5 hold. As $n \to \infty$, $\sqrt{n}\left(\tilde{\theta}_n - \theta^*\right)$ converges in distribution to a zero-mean Gaussian random vector with covariance matrix $\mathbf{C}^{\Theta *}$.

Theorem T3 characterizes the asymptotic distribution of the parameter estimates for the purposes of calculating standard errors of parameter estimates, confidence intervals, and hypothesis testing. For example, the standard error of the ith element of $\tilde{\theta}_n$, σ_i, is defined by the formula: $\sigma_i = (C_{ii}^{\Theta *}/n)^{1/2}$ where $C_{ii}^{\Theta *}$ is the ith on-diagonal element of $\mathbf{C}^{\Theta *}$.

In practice, $\mathbf{C}^{\Theta *}$ is not directly observable and must be estimated. The following theorem provides a mechanisms for estimating $\mathbf{C}^{\Theta *}$ using the *sample sandwich covariance matrix estimator*

$$\tilde{\mathbf{C}}_n^{\Theta} \equiv [\tilde{\mathbf{A}}_n^{\Theta}]^{-1} \tilde{\mathbf{B}}_n^{\Theta} [\tilde{\mathbf{A}}_n^{\Theta}]^{-1}$$

(when it exists).

Theorem T4: Hessian and Covariance Matrix Estimation. Assume A1, A2, and A3 hold. Then as $n \to \infty$, $\mathbf{A}_n^{\Theta}(\theta_n) \to \mathbf{A}^{\Theta *}$ w.p.1, and $\mathbf{B}_n^{\Theta}(\theta_n) \to \mathbf{B}^{\Theta *}$ w.p.1. In addition, if A4 holds, then $\mathbf{C}_n^{\Theta}(\theta_n) \to \mathbf{C}^{\Theta *}$ w.p.1.

The asymptotic covariance matrix $\mathbf{C}^{\Theta *}$ may thus be estimated using $\mathbf{C}_n^{\Theta}(\theta_n)$ provided that A4 and A5 hold. In order to check that A4 and A5 hold, Theorem T4 provides a mechanism for estimating $\mathbf{A}^{\Theta *}$ and $\mathbf{B}^{\Theta *}$ by using $\mathbf{A}_n^{\Theta}(\theta_n)$ and $\mathbf{B}_n^{\Theta}(\theta_n)$ respectively.

1.5 Model Selection

An important problem in model development is the model selection problem. Here, the researcher has a relatively small number of theoretically-motivated models of the same empirical phenomena and wishes to determine which of these models appears to provide the best possible account of the observed data. This concept is now formalized within the context of the framework developed in this article.

Let l^{Θ} be the KDC MAP Objective function for model $\mathcal{M}^{\Theta}$ whose unique global minimum is θ^*. Let l^{Ψ} be the KDC MAP Objective function for model $\mathcal{M}^{\Psi}$ whose unique global minimum is ψ^*. The objective of the *KDC model selection problem* is defined as follows.

- Choose model $\mathcal{M}^{\Theta}$ if $l^{\Theta}(\theta^*) < l^{\Psi}(\psi^*)$.
- Choose model $\mathcal{M}^{\Psi}$ if $l^{\Psi}(\psi^*) < l^{\Theta}(\theta^*)$.
- Choose both models $\mathcal{M}^{\Theta}$ and $\mathcal{M}^{\Psi}$ if $l^{\Psi}(\psi^*) = l^{\Theta}(\theta^*)$.

In practice, the researcher does not have knowledge of the values of $l^{\Theta}(\theta^*)$ and $l^{\Psi}(\psi^*)$ and so these quantities must be estimated from the models $\mathcal{M}^{\Theta}$ and $\mathcal{M}^{\Psi}$ and the data sample. An estimator of l^{Θ}, $\hat{l}_n^{\Theta} \equiv \tilde{l}_n^{\Theta} + k^{\Theta}$, is called a *model selection criterion* where the additional *model selection penalty term* term $k_n^{\Theta} : \Theta \times \Omega^{\tau} \to \mathcal{R}$ is introduced to improve estimation performance for finite samples in a particular way. Note that $k_n^{\Theta} \equiv 0$ implies $\hat{l}_n = \tilde{l}_n$ which yields the *log-likelihood* (also known as the *cross-entropy* or *divergence*) model selection criterion. When k_n^{Θ} is not equal to zero, then it is assumed to converge to zero at a sufficiently fast rate as specified in the following assumption.

Assumption A5. Model Selection Penalty Term. Assume Assumption A1 holds. Assume

$$\sqrt{n} k_n^{\Theta}(\theta_n; \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_n) \to 0$$

in probability as $n \to \infty$.

Assumption A5 requires that the model selection penalty term converges to zero as the sample size n approaches infinity at a particular rate. This constraint upon the convergence rate is not particularly restrictive since many important model selection penalty terms satisfy Assumption $A5$. For example, choosing $k_n^{\Theta} = 0$ results in a *log-likelihood criterion* as previously noted. Let q be the number of free parameters in model $\mathcal{M}^{\Theta}$. Choosing $k_n^{\Theta} = q/n$ results in the well-known *Akaike Information Criterion (AIC)* which yields a model selection criterion for providing a better finite-sample unbiased estimate of $l^{\Theta}(\theta^*)$ under the assumption that the model is correctly specified (Akaike, 1973; Linhart and Zucchini, 1986). Choosing

$$k_n^{\Theta} = trace([\mathbf{A}_n^{\Theta}(\theta_n)]^{-1}(\theta_n)\mathbf{B}_n^{\Theta}(\theta_n))$$

results in the *Generalized Akaike Information Criterion (GAIC)* which extends the AIC to situations where many forms of model misspecification may be present (Linhart and Zucchini, 1986???). Choosing $k_n^{\Theta} = (q/2)log(n)/n$ results in the well-known *Bayes/Schwarz Information Criterion (BIC/SIC)* which yields a model selection criterion to support the selection of the model which is "most probable". A more robust version of BIC/SIC is provided by the *Generalized Bayes Information Criterion (GBIC)* which is defined by choosing

$$k_n^{\Theta} = (1/2n)log\left(det\mathbf{A}_n^{\Theta}(\theta_n)\right) + (q/2n)log(n/2\pi).$$

(Djuric, 1998; also see Shun and McCullagh, 1995). All of the above model selection penalty term choices satisfy Assumption A5 and are valid provided that Assumptions A1, A2, A3, and A4 hold.

Given a particular model selection criterion, a KDC model selection test (Golden, 2003; also see Vuong, 1989; Rivers and Vuong, 2002) may be used to test the null hypothesis that: $H_0 : l^{\Theta}(\theta^*) = l^{\Psi}(\psi^*)$. To achieve this objective, however, the following quantities need to be defined.

Let

$$\mathbf{B}_n^{\Theta,\Psi} \equiv n^{-1} \sum_{i=1}^{n} \sum_{j \in J_{i,\tau}} \mathbf{g}_i^{\Theta} (\mathbf{g}_j^{\Psi})^T.$$

Let

$$\mathbf{B}_n \equiv \begin{bmatrix} \mathbf{B}_n^{\Theta} & \mathbf{B}_n^{\Theta,\Psi} \\ \mathbf{B}_n^{\Psi,\Theta} & \mathbf{B}_n^{\Psi} \end{bmatrix}. \tag{4}$$

Let

$$\mathbf{A}_n \equiv \begin{bmatrix} \mathbf{B}\mathbf{A}_n^{\Theta} & \mathbf{A}_n^{\Theta,\Psi} \\ \mathbf{A}_n^{\Psi,\Theta} & \mathbf{A}_n^{\Psi} \end{bmatrix}. \tag{5}$$

Let

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

where $\mathbf{A}_n$ and $\mathbf{B}_n$ are defined as in (4) and (5).

Let $c_t^{\Theta} \equiv c^{\Theta}(\cdot; \mathbf{f}_t, \ldots, \mathbf{f}_{t-\tau})$ and let $c_t^{\Psi} \equiv c^{\Psi}(\cdot; \mathbf{f}_t, \ldots, \mathbf{f}_{t-\tau})$. Let the *discrepancy variance function*, $(\sigma_Z)^2$

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

Let the discrepancy autocorrelation coefficient function $r_n : \Psi \times \Theta \times \Omega^{\tau+1} \rightarrow \mathcal{R}$ be defined such that:

$$r_n \equiv \frac{\left[\sum_{i=1}^{n} \sum_{j \in J_{i,\tau}, j \neq i} (c_i^{\Theta} - c_i^{\Psi})(c_j^{\Theta} - c_j^{\Psi})\right]}{2\tau \sum_{i=1}^{n} \left(c_i^{\Theta} - c_j^{\Psi}\right)^2}. \tag{6}$$

Assumption A6: KDC Model Selection Assumptions. A6(i): Assume $E\{\mathbf{B}_n\}$ in (4) evaluated at (θ^*, ψ^*) is positive definite. A6(ii): Assume $E\{\mathbf{A}_n\}$ in (5) evaluated at (θ^*, ψ^*) is positive definite. A6(iii): Assume $E\{r_n\}(\theta^*, \psi^*) \neq -1/(2\tau)$ where r_n is defined as in (6).

Given the above quantities, the KDC model selection test may be formulated following the approach of Golden (2003; also see Vuong, 1989; Rivers and Vuong, 2002).

- **Step 1.** Compute the matrix $\mathbf{R}_n$. Let λ_n be a vector whose ith element is the square of the ith eigenvalue of $\mathbf{R}_n$, $i = 1, \ldots, (p+q)$. Compute the probability, p^1, that a weighted chi-square random variable with weight vector equal to λ_n exceeds $n(\sigma^*_{Z_n})^2$. If $p^1 < \alpha$, then go to Step 2. Otherwise, accept $H_0 : l^{\Theta}(\theta^*) = l^{\Psi}(\psi^*)$.
- **Step 2.** Compute the probability, p^2, that a zero-mean Gaussian random variable with unit variance exceeds Z_n. If $p^2 < \alpha$, then reject $H_0 : l^{\Theta}(\theta^*) = l^{\Psi}(\psi^*)$. Otherwise, accept $H_0 : l^{\Theta}(\theta^*) = l^{\Psi}(\psi^*)$.

Assumptions A1, A2, A3, and A6 in conjunction with the theorems provided in Golden (2003) may then be used to establish that the Type 1 error probability associated with the above procedure is asymptotically bounded by significance level α and the Type 2 error probability approaches zero asymptotically (see Golden, 2003, for additional details). Also note that A6 implies A4 and A5 will hold for both models $\mathcal{M}^{\Theta}$ and $\mathcal{M}^{\Psi}$.

1.6 Hypothesis Testing

Theorem T5: Wald Hypothesis Test. Assume A1, A2, A3, A4, and A5 hold. Let $H_0 : \mathbf{S}\theta^* = \mathbf{s}$ where $\mathbf{S} \in \mathcal{R}^{k \times q}$ and $\mathbf{s} \in \mathcal{R}^k$. If H_0 is true, then

$$\tilde{\mathcal{W}}_n \equiv n \left(\mathbf{S}\tilde{\theta}_n - \mathbf{s}\right)^T [\tilde{\mathbf{C}}_n]^{-1} \left(\mathbf{S}\tilde{\theta}_n - \mathbf{s}\right)$$

converges in distribution to a chi-square random variable with k degrees of freedom. If H_0 is false, then $\mathcal{W}_n \to \infty$ w.p.1.

Theorem T5 may be used for *within-group hypothesis testing* applications. For example, consider a KDC probability model consisting of the digraphs $\{\mathcal{D}^1, \ldots, \mathcal{D}^q\}$ with corresponding parameters $\theta_1, \ldots, \theta_q$. Theorem T3 may be used to test null hypotheses such as $H_0 : \theta_k = 0$ or $H_0 : \theta_j = \theta_k$ $(j \neq k)$ through appropriate choices of the selection matrix $\mathbf{S}$ and selection vector $\mathbf{s}$.

Theorem T5 may also be used for *between-group hypothesis testing* applications. Consider a situation involving two data sets which are independent random samples from two DGPs respectively. The two DGPs may be either distinct or identical. Let $\tilde{\theta}^1_n$ denote a q-dimensional parameter estimates obtained from estimating a KDC probability model's parameters using a data sample from DGP 1. Let the covariance matrix of $\tilde{\theta}^1_n$ be denoted by $\tilde{\mathbf{C}}^1_n$. Let $\tilde{\theta}^2_n$ denote a q-dimensional parameter estimates obtained from estimating the same KDC probability model's parameters using a data sample from DGP 2. Let

the covariance matrix of $\tilde{\theta}_n^1$ be denoted by $\tilde{\mathbf{C}}_n^2$. Let θ^{1*} and θ^{2*} denote the pseudo-true parameters corresponding to the asymptotic limits of $\tilde{\theta}_n^1$ and $\tilde{\theta}_n^2$ respectively (when they exist). Now, by defining, $\tilde{\theta} \equiv [\tilde{\theta}_n^1 \;\; \tilde{\theta}_n^2]$ and

$$\tilde{\mathbf{C}}_n \equiv \begin{bmatrix} \tilde{\mathbf{C}}_n^1 & \mathbf{0}_{q\times q} \\ \mathbf{0}_{q\times q} & \tilde{\mathbf{C}}_n^2 \end{bmatrix},$$

it immediately follows that the Wald test in Theorem T5 may be used to test hypotheses regarding whether or not parameter estimates of a KDC Probability Model estimated for one data sample are significantly different from parameter estimates of that same KDC probability model estimated for a second data sample. For example, an omnibus test of the null hypothesis $H_0 : \theta^{1*} = \theta^{2*}$ may be developed or statistical tests for comparing individual corresponding parameters such as: $H_0 : \theta_k^{1*} = \theta_k^{2*}$ for $k = 1, \ldots, q$ can be developed.

2 References

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

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

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

(4) Golden, R. M. (1996). *Mathematical Methods for Neural Network Analysis and Design*. Cambridge, MA: MIT Press.

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

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

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

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

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

(10) White, H. (1984). *Asymptotic Theory for Econometricians*. San Diego: Academic Press.

(11) White, H. (1994). *Estimation, Inference, and Specification Analysis*. New York: Cambridge University Press.

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