Journal of Mathematical Psychology **44**, 62–91 (2000)
doi:10.1006/jmps.1999.1277, available online at http://www.idealibrary.com on IDEAL®

Akaike's Information Criterion and Recent Developments in Information Complexity

Hamparsum Bozdogan

The University of Tennessee

In this paper we briefly study the basic idea of Akaike's (1973) information criterion (AIC). Then, we present some recent developments on a new entropic or *information complexity* (ICOMP) criterion of Bozdogan (1988a, 1988b, 1990, 1994d, 1996, 1998a, 1998b) for model selection. A rationale for ICOMP as a model selection criterion is that it combines a badness-of-fit term (such as minus twice the maximum log likelihood) with a measure of complexity of a model differently than AIC, or its variants, by taking into account the interdependencies of the parameter estimates as well as the dependencies of the model residuals. We operationalize the general form of ICOMP based on the quantification of the concept of overall model complexity in terms of the estimated *inverse-Fisher information matrix*. This approach results in an approximation to the sum of two *Kullback–Leibler distances*. Using the correlational form of the complexity, we further provide yet another form of ICOMP to take into account the interdependencies (i.e., *correlations*) among the parameter estimates of the model. Later, we illustrate the practical utility and the importance of this new model selection criterion by providing several real as well as Monte Carlo simulation examples and compare its performance against AIC, or its variants.

INTRODUCTION AND PURPOSE

In recent years, the statistical literature has placed more and more emphasis on model evaluation criteria. The necessity of introducing the concept of model evaluation has been recognized as one of the important technical areas, and the problem is posed on the choice of the best approximating model among a class of competing models by a suitable model evaluation criteria given a data set. Model evaluation criteria are figures of merit, or performance measures, for competing models. In this paper, we shall briefly study the basic underlying idea of Akaike's (1973) information criterion AIC. Then, we introduce a new information-theoretic measure of complexity criterion called ICOMP of Bozdogan (1987b, 1988a, 1988b, 1990, 1994d, 1996) as a decision rule for model selection and evaluation.

Recently, based on Akaike's (1973) original AIC, many model-selection procedures which take the form of a penalized likelihood (a negative log likelihood plus

0022-2496/00 $35.00

a penalty term) have been proposed (Sclove, 1987). For example, for AIC this form is given by

$$\mathrm{AIC} = -2 \log L(\hat{\theta}) + 2k, \tag{1}$$

where $L(\hat{\theta})$ is the maximized likelihood function, and k is the number of free parameters in the model. The model with minimum AIC value is chosen as the best model to fit the data.

In AIC, the compromise takes place between the maximized log likelihood, i.e., $-2 \log L(\hat{\theta})$ (the lack of fit component) and k, the number of free parameters estimated within the model (the penalty component) which is a measure of complexity or the compensation for the bias in the lack of fit when the maximum likelihood estimators are used. In using AIC, according to Akaike (1987, p. 319), the accuracy of parameter estimates is measured by a universal criterion, namely

$$Accuracy Measure = \mathscr{E}[\, log\ likelihood\ of\ the\ fitted\ ], \tag{2}$$

where $\mathscr{E}$ denotes the expectation, since AIC is an unbiased estimator of minus twice the expected log likelihood.

The development of ICOMP has been motivated in part by AIC, and in part by *information complexity concepts* and *indices*. In contrast to AIC, we base the new procedure ICOMP on the *structural complexity* of an element or set of random vectors via a generalization of the *information-based covariance complexity index* of van Emden (1971). For a general multivariate linear or nonlinear model defined by

$$Statistical\ Model = Signal + Noise \tag{3}$$

ICOMP is designed to estimate a loss function

$$Loss = Lack\ of\ fit + Lack\ of\ parsimony + \mathrm{Pr}\ ofusion\ of\ Complexity \tag{4}$$

in several ways. This is achieved by using the additivity property of information theory and the entropic developments in Rissanen (1976) in his *final estimation criterion* (*FEC*) in estimation and model identification problems, as well as Akaike's (1973) AIC and its analytical extensions in Bozdogan (1987a). In the loss function (4), by the third term, *profusion of complexity*, we mean the *interdependencies* or the *correlations among the parameter estimates* and the *random error* term of a model.

We propose a general approach to ICOMP. This approach, referred to as ICOMP(IFIM), exploits the well-known asymptotic optimality properties of the *maximum likelihood estimators* (MLEs), and use the information-based complexity of the *inverse-Fisher information matrix* (IFIM) of a model. This is known as the *Cramér–Rao lower bound* (CRLB) *matrix*. This approach results in an approximation to the sum of two *Kullback–Leibler distances*. Using the correlational form of the complexity, we further obtain another form of ICOMP, namely $\mathrm{ICOMP(IFIM)}_R$. This form takes into account the interdependencies (i.e., *correlations*) among the parameter estimates and the model residuals. Later we also give other versions of

ICOMP based on the finite sampling distributions of the parameter estimates which are useful in linear models that take account of complexity of both the parameter structure and the random error structure of the model.

Each approach takes into account the accuracy of the estimated parameters of the model and gives us the flexibility to investigate the influence of different error covariance structures on the accuracy of the parameter estimates. Furthermore, each formulation of ICOMP has the attractive feature of implicitly adjusting for the *number of parameters*, the *sample size*, and *controlling the risks of both insufficient and overparameterized models*.

Comparing with AIC, in ICOMP, complexity is viewed not as the number of parameters in the model, but as the *degree of interdependence* among the components of the model. In the literature, several authors (e.g., Rissanen, 1976) have questioned whether the penalty term $2k$ in AIC is a sufficient penalty term in order to prevent overspecialization and unnecessary complexity by the chosen model. In ICOMP by defining complexity in terms of the *degree of interdependence* among the components of the model, our objective is to provide a more judicious penalty term than AIC and other AIC-*type criteria*, since counting and penalizing the number of parameters in a model is necessary but by no means sufficient. Model complexity in statistics depends intrinsically on many factors other than the model dimension, such as the several forms of *parameter redundancy*, *parameter stability*, random *error structure* of the model, and the *linearity and nonlinearity of the parameters* of the model, to mention a few.

Therefore, in ICOMP, in addition to *lack of fit*, the *lack of parsimony* and the *profusion of complexity* are data-adaptively adjusted by the entropic complexity of the *estimated* IFIM across the competing alternative models as the parameter spaces of these models are constrained in the model fitting process data-adaptively.

The basic approach is that a model with minimum ICOMP is chosen to be the best model among all competing models. Other things equal, the best model is the one which achieves the most satisfactory compromise between the accuracy of the estimated model parameters and the interactions of the residuals. The general principle is that, for a given level of accuracy, a simpler model (i.e., one with a *small covariance matrix of the parameter estimates* and a *small residual covariance matrix*) is preferable to a more complex one. Here small is used in the sense of *minimum variance*.

Hence, the main purpose of this paper is to develop and present *information-theoretic ideas of a measure of overall model complexity* in *statistical estimation* to help provide new approaches relevant to *statistical inference and inquiry*.

Due to space limitations, the detailed proofs and derivations will not be shown in this paper, except that we will only show why ICOMP(IFIM) is an approximation to the sum of two *Kullback–Leibler distances* and explain briefly the theory behind this general approach. For more details, we will refer the reader to Bozdogan (1987a, 1998a, 1998b), Bozdogan and Bearse (2000), and Bozdogan and Haughton (1998).

We illustrate the practical utility and the importance of this new model selection criterion by providing several real examples as well as a Monte Carlo simulation example and compare its performance against AIC and AIC-*type criteria*.

AKAIKE'S INFORMATION CRITERION (AIC)

In statistical model evaluation in choosing the best approximating model from finite samples, we encounter two types of errors: (i) *error caused by modeling*, and (ii) *error done by estimation of the parameter vector* θ, where we also encounter what is called the *estimation error*, namely the *bias* and *variance*. Let R denote the overall risk, $R(M)$ denote the risk of modeling, and $R(E)$ denote the risk of estimation. Then, we can define

$$\underbrace{\textit{Overall Risk}}_{R} = \underbrace{\textit{Risk of modeling}}_{R(M)} + \underbrace{\textit{Risk of estimation.}}_{R(E)} \tag{5}$$

Generally, when we are using model-selection criteria, we fit models under a specified parametric probability distribution of the model. Often, during the course of the analysis of data, we may discover that the particular form of the specified parametric probability model may not be the appropriate distribution for the data at hand. In such a case, we encounter a risk of modeling in terms of the correct specification of the distribution of the model. As is well known, the correct specification of the probability model is sufficient, but by no means a necessary condition. *Risk of estimation* is encountered when we estimate the true parameter vector in the restricted parameter space of the model. This risk is called the *variance component in estimation*. When the true parameter vector is excluded from the restricted parameter space of the model, then a *bias* is caused. Another way to interpret the variance and bias in estimation is as follows. *The variance can be interpreted as a penalty for the size of the admitted parameter space of the model*, and *bias is a penalty for the distance between the reduced or restricted parameter space and the true parameter vector of the model*. In model selection, our goal, then, is to minimize the overall risk R. In this sense, model-selection criteria are the estimators of the overall risk of a model under the maximum likelihood estimation and are called figures of merit. Akaike, in a very important sequence of papers, including Akaike (1973), (1974), and (1981), pioneered for us the field of statistical data modeling and statistical model identification or evaluation. The school of such activity is now called the *Akaike school*.

Kullback–Leibler Information as a Measure of Goodness of Fit

The rationale of Akaike's concept of choosing the best approximating model from finite samples can be formulated as *maximizing entropy*, or equivalently *minimizing Kullback–Leibler* (*KL*) (1951) *information*.

Consider a probability density function $f(\mathbf{x}\,|\,\theta^*)$ of a continuous random variable $\mathbf{x}$ of interest, and let $f(\mathbf{x}\,|\,\theta) \equiv g(\mathbf{x}\,|\,\theta)$ be the density function that specifies the model or is an approximation to $f(\mathbf{x}\,|\,\theta^*)$ with a given parameter vector $\theta \equiv \theta_k = (\theta_1, \theta_2, ..., \theta_k) \in \Re^k$. We will measure the *closeness*, or the *goodness-of-fit*, of $f(\mathbf{x}\,|\,\theta^*)$ with respect to the model $f(\mathbf{x}\,|\,\theta) \equiv g(\mathbf{x}\,|\,\theta)$ by the generalized entropy (B) of Boltzmann (1877) or Kullback–Leibler (1951) information (I).

We minimize the negentropy or the KL information,

$$\begin{aligned} I(\theta^*; \theta) &= -B(\theta^*; \theta) \\ &= E_x[\log f(\mathbf{X} \mid \theta^*) - \log g(\mathbf{x} \mid \theta)] \\ &= \int_{\mathfrak{R}} f(\mathbf{x} \mid \theta^*) \log f(\mathbf{x} \mid \theta^*)\, d\mathbf{x} - \int_{\mathfrak{R}} f(\mathbf{x} \mid \theta^*) \log g(\mathbf{x} \mid \theta)\, d\mathbf{x} \\ &= H(\theta^*; \theta^*) - H(\theta^*; \theta), \end{aligned} \tag{6}$$

where $H(\theta^*; \theta^*) \equiv H(\theta^*)$ is the usual negative entropy which is constant for a given $f(\mathbf{x} \mid \theta^*)$, and where $H(\theta^*; \theta)$ is the cross-entropy which determines the goodness of fit of $g(\mathbf{x} \mid \theta)$ to $f(\mathbf{x} \mid \theta^*)$. In (6), and throughout the paper *log* denotes the *natural logarithm.*

The quantity $H(\theta^*; \theta)$ plays a crucial role in the development of AIC and is of basic importance in statistical information theory. The analytic properties of $I(\theta^*; \theta)$ are extensively discussed by Kullback (1968). Here we list some of the important ones.

- $I(\theta^*; \theta) > 0$ whenever $f(\mathbf{x} \mid \theta^*) \neq g(\mathbf{x} \mid \theta)$,
- $I(\theta^*; \theta) = 0$, if and only if $f(\mathbf{x} \mid \theta^*) = g(\mathbf{x} \mid \theta)$ almost everywhere (a.e.) in the possible range of $\mathbf{x}$, when the model is essentially true,
- *if* $X_1, X_2, ..., X_n$ are independent and identically distributed (i.i.d.) random variables, then the KL information for the whole sample is $I_n(\theta^*; \theta) = nI(\theta^*; \theta)$.

This last property says that if the random variables are independent, the KL information is additive. We note that KL information quantity is perhaps the most general of all information measures in the sense of being derivable from minimal assumptions and it represents a relative measure of distance. But, we remark that KL information is not a metric on the space of probability densities since it does not satisfy the usual triangle inequality and it is not symmetric. Nevertheless, the use of KL information as a loss function is justified since it is a measure of *closeness*, or the *goodness-of-fit*, and can be interpreted as the *mean information for discrimination in favor of* $f(\mathbf{x} \mid \theta^*)$ against $g(\mathbf{x} \mid \theta)$, and it will induce a Riemannian metric on a parameter manifold under suitable conditions (Balasubramanian, 1996, p. 8). As it stands, the KL information quantity in (6) is not directly observable or estimable since it depends on the true distribution and consequently on the unknown true and model parameters. Therefore, maximization of the mean log likelihood is carried out, and asymptotically an unbiased estimator of the mean expected log likelihood is searched by correcting the bias of the observed mean log likelihood.

AIC as a Bias Correcting Criterion

Since $H(\theta^*; \theta^*) \equiv H(\theta^*)$ is a constant in (6), we only have to estimate the cross-entropy in KL information. That is, we need only to estimate

$$H(\theta^*; \theta) = E_x[\log f(\mathbf{X} \mid \theta)] = \int_{\mathfrak{R}} f(\mathbf{x} \mid \theta^*) \log g(\mathbf{x} \mid \theta)\, d\mathbf{x} \tag{7}$$

which is the *expected log likelihood of the model's pdf* $g(\mathbf{x} \mid \theta)$ with respect to $f(\mathbf{x} \mid \theta^*)$.

Assuming that a sample of n observations $\mathbf{x} = (x_1, x_2, ..., x_n)$ is used to provide an estimate $\hat{\theta} \equiv \hat{\theta}(\mathbf{x})$ of θ, maximizing the average or mean log likelihood is asymptotically equivalent to minimizing the *KL information* $\hat{I}(\theta^*; \theta)$, where $\hat{I}(\theta^*; \theta)$ denotes an estimator of $I(\theta^*; \theta)$ (e.g., *ML estimator*). Hence, asymptotically the maximum of

$$\frac{1}{n} \log L(\mathbf{x} \mid \hat{\theta}) = \int_{\mathfrak{R}} f(\mathbf{x}) \log f(\mathbf{x})\, d\mathbf{x} = \hat{H}(\theta^*; \theta^*) \tag{8}$$

occurs when $\hat{I}(\theta^*; \theta) = 0$ almost everywhere. This happens if and only if $f(\mathbf{x} \mid \theta^*) = g(\mathbf{x} \mid \theta)$ almost everywhere in the possible range of $\mathbf{x}$, when the model is essentially true, and that the estimated entropy $\hat{H}(\theta^*; \theta^*)$ plays the role of an asymptote. Hence, (8) is an important relationship between cross-entropy and the likelihood which makes the use of the likelihood function clear. Since the *average log likelihood* is an estimate of cross-entropy, when it is maximized, then the estimated cross-entropy $\hat{H}(\theta^*; \hat{\theta})$ is minimized. The smaller $\hat{H}(\theta^*; \hat{\theta})$ is, the better the model approximates the true model in terms of entropy. Hence, it is natural to compare models by using the *maximized average log likelihood.* However, it is well known that such a method produces estimation bias (or overestimation) when comparing models with different sizes based on a finite number of observations, because the same set of observations is used to estimate the parameters of the model which in turn estimates $H(\theta^*; \hat{\theta})$.

Indeed in defining AIC, Akaike (1973, 1974) has exactly this consideration of correction of the estimation bias by penalizing extra parameters when the MLEs are used in estimating the expected log likelihood. Following Kitagawa and Gersch (1996), Konishi and Kitagawa (1996), and Konishi (1998), we give the *bias between the average of the maximized log likelihood* and *the expected maximized log likelihood* as

$$b = Bias = E_G \left[\frac{1}{n} \sum_{i=1}^{n} \log f(x_i \mid \hat{\theta}) - \int_{\mathfrak{R}} \log f(\mathbf{x} \mid \hat{\theta})\, dG(\mathbf{x}) \right], \tag{9}$$

where the expectation is taken over the true distribution $G = \prod_{i=1}^{n} dG(x_i)$. Note that this bias b is not equal to zero, since $\hat{\theta}$ is a function of $\mathbf{x}$, and the two instances of $\hat{\theta}$ are tied to different integration variables in (9). Therefore, the estimated bias $\hat{b}$ is generally obtained as an asymptotic bias and as an approximation to b. So, if the bias b can be estimated by appropriate procedures, then we can define an information criterion based on the bias corrected log likelihood. Hence, we have the following.

DEFINITION 1. A bias corrected information criterion (BCIC) is defined by

$$\begin{aligned} \text{BCIC} &= -2n \left\{ \frac{1}{n} \sum_{i=1}^{n} \log f(x_i \mid \hat{\theta}) - b \right\} \\ &= -2 \sum_{i=1}^{n} \log f(x_i \mid \hat{\theta}) + 2nb \\ &= -2 \log L(\hat{\theta}) + 2nb. \end{aligned} \tag{10}$$

This definition is important since a large area of research in statistics deals with the finite sample properties of linear and nonlinear (MLEs) in statistical modeling. For example, in nonlinear models, say, in nonlinear regression, these estimators are generally biased of the true model parameter values. In fact, the bias, in general, is of order $O(n^{-1})$, where n is the sample size. Therefore, it will be very useful in different model specific settings to have exact formulas available for calculating the bias, rather than ignoring the bias in practice with a weak justifcation that it is negligible when compared to the standard errors of the parameter estimates. For example, in the usual multiple regression case the exact bias b of the log likelihood is calculated as

$$b = Bias = E_G\left[\frac{1}{n}\sum_{i=1}^{n}\log f(x_i\,|\,\hat{\theta}) - \int_{\mathfrak{R}}\log f(\mathbf{x}\,|\,\hat{\theta})\,dG(\mathbf{x})\right] = \frac{n(k+1)}{n-k-2}. \quad (11)$$

In general, the bias expression given in (9) needs to be derived under other model specific settings. In some cases it is difficult, if not impossible, to derive the expression in (9) in closed form. In cases when we do not have exact formulas available for calculating the bias, then one can utilize the bootstrap bias in BCIC in (10). This idea at the outset is appealing, but it is not practical in the sense that such an approach becomes too costly to compute. It becomes very computer intensive and does not provide us a simple yet practical just-in-time model selection criterion.

Akaike (1973), in deriving his AIC, did not ignore this bias in estimation, but went to asymptotics too quickly and under the assumptions that: (i) *the parametric model is estimated by the method of maximum likelihood*, and (ii) *the specified parametric family of p.d.f.'s contains the true distribution* (i.e., $g(\mathbf{x}) = f(\mathbf{x}\,|\,\theta^*)$ for some $\theta^* \in \Theta$, the parameter space), he approximated this bias to be

$$b = Bias = \frac{k}{n}. \quad (12)$$

Hence:

DEFINITION 2. Akaike's (1973) information criterion (AIC) is defined by

$$\begin{aligned} \mathrm{AIC} &= -2\sum_{i=1}^{n}\log f(x_i\,|\,\hat{\theta}) + 2k \\ &= -2\log L(\hat{\theta}) + 2k. \end{aligned} \quad (13)$$

Note that the bias in AIC is approximated by the number of parameters which is constant and has no variability. If we use the result in (11) for the normal multiple regression model then for small samples, we can define the finite sample corrected AIC, namely, AIC_c, originally proposed by Sugiura (1978) and later used by Hurvich and Tsai (1989) given by

$$\mathrm{AIC}_c = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2\frac{n(k+1)}{n-k-2}, \quad (14)$$

where $\hat{\sigma}^2$ is the variance of the residuals, and k is the number of predictors in the regression model. This implies that AIC_c in (14) in the normal regression model achieves bias reduction exactly under the assumption that the true distribution belongs to the specified parametric family (Konishi 1998, p. 24).

Following Takeuchi (1976) and Shibata (1989), in practice we are interested in choosing a model among several different parametric models or non-nested models. In such cases, we can relax the assumption of Akaike. That is, we *assume* that the *true distribution does not belong to the specified parametric family of p.d.f.'s*. In other words, if the parameter vector θ of the distribution is unknown and is estimated by the maximum likelihood method, then it is not any longer true that the *average of the maximized log likelihood* converges to the *expected value of the parameterized log likelihood*. That is,

$$\frac{1}{n}\log L(\mathbf{x}\,|\,\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\log f(x_i\,|\,\hat{\theta}) \not\to E_x[\log f(\mathbf{X}\,|\,\hat{\theta})]. \tag{15}$$

In this case, the bias, b, between the *average of the maximized log likelihood* and the *expected maximized log likelihood* is given by

$$b = Bias = E_G\left[\frac{1}{n}\sum_{i=1}^{n}\log f(x_i\,|\,\hat{\theta}) - \int_{\mathfrak{R}}\log f(\mathbf{x}\,|\,\hat{\theta})\,dG(\mathbf{x})\right] = \frac{1}{n}\,tr(\mathscr{F}^{-1}\mathscr{R}) + O(n^{-1}), \tag{16}$$

where $\mathscr{F}$ is the *inverse Fisher information matrix in inner product* or *Hessian form*, and $\mathscr{R}$ is the *outer product form of the Fisher information matrix* both with dimension $(k \times k)$. We note that $tr(\mathscr{F}^{-1}\mathscr{R})$ is the well-known Lagrange-multiplier test statistic. See, for example, Takeuchi (1976), Hosking (1980), and Shibata (1989). Thus, we have:

DEFINITION 3. Generalized Akaike's (1973) information criterion (GAIC) is defined by

$$\begin{aligned}\mathrm{GAIC} &= -2\sum_{i=1}^{n}\log f(x_i\,|\,\hat{\theta}) + 2tr(\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}) \\ &= -2\log L(\hat{\theta}) + 2tr(\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}).\end{aligned} \tag{17}$$

In the literature of model selection, GAIC is also known as Takeuchi's (1976) information criterion (TIC), or AIC_T.

When the probability model is correctly specified and certain regularity conditions hold from White (1982, p. 6) we have:

THEOREM 1 (Information Matrix Equivalence Test). *If the fitted probability model* $f(\mathbf{x}) \equiv f(\mathbf{x}\,|\,\theta^*)$ *for* $\theta^* \in \Theta$, *then* $\theta^* = \theta_k^*$ *and* $\mathscr{F}(\theta_k^*) = \mathscr{R}(\theta_k^*)$, *so that*

$$\begin{aligned}Cov(\theta_k^*) &= \mathscr{F}^{-1}(\theta_k^*)\,\mathscr{R}(\theta_k^*)\,\mathscr{F}^{-1}(\theta_k^*) \\ &= \mathscr{F}^{-1}(\theta_k^*) = \mathscr{R}^{-1}(\theta_k^*).\end{aligned} \tag{18}$$

The basic idea underlying the *Information Matrix Equivalence Test* (*IMET*) is that it relies on the equality between the two forms of the Fisher information matrix which are useful to check the misspecification of a model. So, if $\mathscr{F}(\theta_k^*) = \mathscr{R}(\theta_k^*)$, then the bias reduces to

$$b = Bias = \frac{1}{n} tr(\mathscr{F}^{-1}\mathscr{R}) + O(n^{-2}) = \frac{1}{n} tr(I_k) + O(n^{-2})$$

$$= \frac{1}{n} k + O(n^{-2}), \tag{19}$$

which gives AIC in (13) as a special case. For more details on these, we refer the readers to Bozdogan (1998a, 1998b), Kitagawa and Gersch (1996), Konishi and Kitagawa (1996), and Konishi (1998).

We interpret the above results as follows. For AIC in (13), the first term provides us with *a measure of bias* or *model inaccuaracy* (*badness of fit* or *lack of fit*) when the maximum likelihood estimators of the parameters of the model are used. The second term serves *a penalty* for the increased unreliability or compensation for the bias in the first term when additional free parameters are included in the model. Thus, when there are several competing models, the parameters within the models are estimated by the method of maximum likelihood and the values of AICs are computed and compared to find a model with the *minimum value of AIC*. This procedure is called the *minimum AIC procedure* and the model with the minimum AIC is called the *minimum AIC estimate* (*MAICE*) and is chosen to be the best model. As is well known, as the sample size gets large, the first term of AIC increases but the penalty term $2k$ does not, since it is fixed. This means that the penalty term has little effect if the sample size n is large. For this reason, objections have been raised that minimizing AIC does not produce an asymptotically consistent estimate of model order (Bhansali and Downham, 1977; Schwarz, 1978; Woodroofe, 1982; and others). However, we should note that consistency is an asymptotic property, and any real problem has a finite sample size n (Sclove, 1987). Therefore, this charge on AIC should not be exaggerated (Hannan, 1986; Forster, 1999). In general, the application of AIC emphasizes the comparison of goodness of fit of the competing models while taking into account the principle of parsimony. Certainly AIC provides an answer to the question of how much improvement in fit an additional parameter should make before it is included in the model and on what scale that improvement should be measured.

Of course, important fundamental work like this answers some questions and raises many others. Without violating Akaike's main setup, using the device of effective degrees of freedom (df), i.e., correcting for the df (see, e.g., Cox, 1984), Bozdogan (1987a) extended AIC in several ways to improve and modify AIC to make it consistent. For more on the *general theory*, *consistency*, *inferential error rates*, and many applications of AIC, *consistent Akaike's information criterion* (*CAIC*), and *consistent AIC with Fisher Information* (*CAICF*), we refer the readers to Bozdogan (1987a, 1994a–1994c).

We interpret GAIC similar to AIC with the exception that the penalty term is now $2tr(\hat{\mathscr{F}}^{-1}\hat{\mathscr{R}})$. We use GAIC when we are in doubt of whether the class of potential models are correctly specified or not by testing the relationship between the two forms of the estimated information matrices. In other words, we use GAIC to guard ourselves against *misspecifying a model*. Furthermore, one should use GAIC to guard against *high skewness* and *kurtosis* in the data. In the case of high skewness and kurtosis, the penalty in AIC is not adequate to compensate for the bias in the maximum likelihood estimates of the parameters of a model. Also, in equivalent models, that is, models with equivalent structures having the same number of parameters, AIC will not be able to distinguish one model from the other due to the fact that $2k$ is fixed and has no variability. Therefore, by taking the higher order bias correction terms into account, we may obtain more refined criteria. However, as we discussed above, in some cases it is difficult, if not impossible, to obtain the bias of a log likelihood in closed form.

This brings us to another way to derive information theoretic model selection criteria, namely, to ICOMP criterion of Bozdogan (1988a, 1988b, 1990, 1994d), to provide a more judicious penalty term and to balance the overfitting and underfitting risks of a model than that of AIC. Indeed, this new approach provides an entropic general *data-oriented* or what we also call a *data-adaptive penalty functional*, which is random and is an improvement over a fixed choice of penalty functional such as in AIC, or its variants.

RECENT DEVELOPMENTS IN INFORMATION COMPLEXITY

In this part of the paper, we introduce a new model selection criterion:

- To measure the fit between multivariate structural models and observed data as an example of the application of the covariance complexity measure.
- To allow the measurement of dependency between the random variables.
- To establish and provide a trade-off between the fit and the interaction of the parameter estimates and the interaction of the residuals of a model via the measure of complexity of their respective covariances.
- To remove from the researcher any need to consider the parameter dimension explicitly, since the bias in AIC is approximated by the number of parameters which is constant and has no variability, and
- To provide a more judicious penalty term than AIC, or AIC-type criteria to balance the overfitting and underfitting risks of a model.

The Concept of Complexity and Complexity of a System

Complexity is a general property of statistical models that is largely independent of the specific content, structure, or probabilistic specification of the models. In the literature, the concept of complexity has been used in many different contexts. In general, there is not a unique definition of complexity in statistics, since the notion is *elusive* according to van Emden (1971, p. 8). Complexity has many faces, and it is defined under many different names such as those of *Kolmogorov complexity* (Cover,

Gacs, and Gray, 1989), *Shannon complexity* (Rissanen, 1989), and *Stochastic complexity* (Rissanen, 1987, 1989) in information theoretic coding theory, to mention a few. For example, Rissanen (1986, 1987, 1989), similar to Kolmogorov (1983), defines complexity in terms of the *shortest code length for the data that can be achieved by the class of models*, and calls it *Stochastic complexity* (SC). The Monash school (e.g., Wallace and Freeman, 1987, Wallace and Dowe, 1993, Baxter, 1996) define complexity in terms of *minimum message length* (MML), which is based on evaluating models according to their ability to compress a message containing the data. In our case, complexity is defined in terms of the following simple system theoretic definition which later motivates a statistically defined measure based on entropy maximization.

DEFINITION 4. Complexity of a system (of any type) is a measure of the degree of interdependency between the whole system and a simple enumerative composition of its subsystems or parts.

We note that this definition of complexity is different from the way it is now frequently used in the literature to mean the number of estimated parameters in a model. For our purposes, the complexity of a model is most naturally described in terms of interactions of the components of the model and the information required to construct the model in a way it is actually defined. Therefore, the notion of complexity can be best explained if we consider the statistical model arising within the context of a real world system. For example, the system can be physical, biological, social, behavioral, economic, etc., to the extent that the system responses are considered to be random.

As we defined in Definition 4, the *complexity of a system* (*of any type*) is a *measure of the degree of interdependency between the whole system and a simple enumerative composition of its subsystems or parts.* Naturally, we are interested in the amount by which the whole system, say, S, is different from the composition of its components. If we let C denote any real-valued measure of complexity of a system S, then $C(S)$ will measure the amount of the difference between the whole system and its decomposed components. Using the information theoretic interpretation, we define this amount to be the discrimination information of the joint distribution of the probability model at hand against the product of its marginal distributions. Discrimination information is equal to zero if the distributions are identical and is positive otherwise (van Emden 1971, p. 25).

Thus, to quantify the concept of complexity in terms of a *scalar index*, we only have to express the interactions in a mathematical definition. We shall accomplish this by appealing to information theory since it possesses several important analytical advantages over the conventional procedures, such as those of additivity and constraining properties, and allowance to measure dependencies.

For more details on the system theoretic definition of complexity as background material, we refer the reader to van Emden (1971, pp. 7 and 8), and Bozdogan (1990).

Initial Definition of Information Theoretic Covariance Complexity

For a random vector, we define the complexity as follows.

DEFINITION 5. The complexity of a random vector is a measure of the interaction or the dependency between its components.

We consider a continuous p-variate distribution with joint density function $f(\mathbf{x}) = f(x_1, x_2, ..., x_p)$ and marginal density functions $f_j(x_j)$, $j = 1, 2, ..., p$. Following Kullback (1968), Harris (1978), Theil and Fiebig (1984), and others, we define the *informational measure of dependence* between random variables $x_1, x_2, ..., x_p$ by

$$I(\mathbf{x}) \equiv I(x_1, x_2, ..., x_p) = E_f\left[\log \frac{f(x_1, x_2, ..., x_p)}{f_1(x_1) f_2(x_2) \cdots f_p(x_p)}\right]$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, ..., x_p) \log \frac{f(x_1, x_2, ..., x_p)}{f_1(x_1) f_2(x_2) \cdots f_p(x_p)} dx_1 \cdots dx_p, \qquad (20)$$

where I is the *Kullback–Leibler* (1951) *information divergence against independence*. $I(\mathbf{x})$ in (20) is a *measure of expected dependency* among the component variables, which is also known as the *expected mutual information* or the *information proper*.

- Property 1. $I(\mathbf{x}) \equiv I(x_1, x_2, ..., x_p) \geqslant 0$, i.e., the expected mutual information is nonnegative.

- Property 2. $f(x_1, x_2, ..., x_p) = f_1(x_1) f_2(x_2) \cdots f_p(x_p)$ for every p-tuple $(x_1, x_2, ..., x_p)$, i.e., if and only if the random variables $x_1, x_2, ..., x_p$ are mutually statistically independent, then the quotient in (20) is equal to unity, and its logarithm is then zero. Hence, $I(\mathbf{x}) \equiv I(x_1, x_2, ..., x_p) = 0$. If it is not zero, this implies a dependency.

We relate the *KL divergence* in (20) to *Shannon's* (1948) *entropy* by the important identity

$$I(\mathbf{x}) \equiv I(x_1, x_2, ..., x_p) = \sum_{j=1}^{p} H(x_j) - H(x_1, x_2, ..., x_p), \qquad (21)$$

where $H(x_j)$ is the marginal entropy, and $H(x_1, x_2, ..., x_p)$ is the global or joint entropy. Watanabe (1985) calls (21) the *strength of structure* and a *measure of interdependence*. We note that (21) is the sum of the interactions in a system with $x_1, x_2, ..., x_p$ as components, which we define to be the entropy complexity of that system. This is also called the *Shannon Complexity* (see Rissanen, 1989). If there exists more interdependency in the structure, we will see that the more markedly the sum of the marginal entropies will be. Consequently, this will dominate the joint entropy. If we wish to extract fewer and more important variables, it will be desirable that they be statistically independent, because the presence of interdependence means redundancy and mutual duplication of information contained in these variables (Watanabe, 1985).

The relation in (21) can easily be generalized to finding the interaction between any subset of variables also. Suppose we consider two sets of random variables $(x_1, x_2, ..., x_p)$ and $(x_{p+1}, ..., x_{p+q})$. For the interaction we have

$$I((x_1, x_2, ..., x_p), (x_{p+1}, ..., x_{p+q}))$$
$$= H(x_1, x_2, ..., x_p) + H(x_{p+1}, ..., x_{p+q}) - H(x_1, x_2, ..., x_{p+q}). \quad (22)$$

Now, to define the information-theoretic measure of complexity of a multivariate distribution, we let $f(\mathbf{x}) \equiv f(x_1, x_2, ..., x_p)$ be a multivariate normal density function with a p-dimensional mean vector μ and $(p \times p)$ positive definite (pd) covariance matrix Σ. After some work, we then find for the total amount of interaction:

$$I(\mathbf{x}) \equiv I(x_1, x_2, ..., x_p) = \sum_{j=1}^{p} H(x_j) - H(x_1, x_2, ..., x_p)$$
$$= \sum_{j=1}^{p} \left[\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\sigma_j^2) + \frac{1}{2} \right] - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{p}{2}. \quad (23)$$

This reduces to

$$C_0(\Sigma) = \tfrac{1}{2} \sum_{j=1}^{p} \log(\sigma_j^2) - \tfrac{1}{2}\log|\Sigma|, \quad (24)$$

where σ_j^2 is the jth diagonal element of Σ and p is the dimension of Σ. Note that $C_0(\Sigma) = 0$ when Σ is a diagonal matrix (i.e., if the variates are linearly independent). $C_0(\Sigma)$ is infinite if any one of the variables may be expressed as a linear function of the others ($|\Sigma| = 0$). If $\theta = (\theta_1, \theta_2, ..., \theta_k)$ is a normal random vector with covariance matrix equal to $\Sigma(\theta)$, then $C_0(\Sigma(\theta))$ is simply the KL distance between the multivariate normal density of θ and the product of the marginal densities of the components of θ. As pointed out by van Emden (1971), C_0 is not an effective measure since it depends upon marginal and joint distributions of the random variable and it is not invariant under orthonormal transformations.

Definition of Maximal Covariance Complexity

Since we defined the complexity as a general property of statistical models, we consider that the general definition of complexity of a covariance matrix Σ should be independent of the coordinates of the original random variables $x_1, x_2, ..., x_p$ associated with the variances σ_j^2, $j = 1, 2, ..., p$. As it is, $C_0(\Sigma)$ in (24) is coordinate dependent. However, to characterize the maximal amount of complexity of Σ, we can relate the general definition of complexity of Σ to the total amount of interaction or $C_0(\Sigma)$ in (24). We do this by recognizing the fact that the maximum of (24) under orthogonal transformation of the coordinate system may reasonably serve as the measure of complexity of Σ. This corresponds to observing the interaction between the variables under the coordinate system that clearly represents it in terms of the measure $I(x_1, x_2, ..., x_p) \equiv C_0(\Sigma)$.

So, to improve on (24), we have the following definition.

DEFINITION 6. A maximal information theoretic measure of complexity of a covariance matrix Σ of a multivariate distribution is

$$C_1(\Sigma) = \max_T C_0(\Sigma)$$
$$= \frac{p}{2}\log\left[\frac{tr(\Sigma)}{p}\right] - \frac{1}{2}\log|\Sigma|, \tag{25}$$

where the maximum is taken over orthonormal transformation T of the overall coordinate systems $x_1, x_2, ..., x_p$.

$C_1(\Sigma)$ in (25) is an upper bound to $C_0(\Sigma)$ in (24), and it measures both inequality among the variances and the contribution of the covariances in Σ (van Emden, 1971, p. 63). Such a measure is very important in constructing model selection criteria to determine the strength of model structures, similarity, dissimilarity, and high-order correlations within the model. $C_1(\Sigma)$ is independent of the coordinate system associated with the variances σ_j^2, $j = 1, 2, ..., p$. Furthermore, if, for example, one of the σ_j^2's is equal to zero, then $C_0(\Sigma)$ in (24) takes the value $\infty - \infty$, which is indeterminate, whereas $C_1(\Sigma)$ in (25) has the value ∞ (infinity) which has a mathematical meaning. Also, $C_1(\Sigma)$ has rather attractive properties. Namely, $C_1(\Sigma)$ is invariant with respect to scalar multiplication and orthonormal transformation. Further, $C_1(\Sigma)$ is a monotonically increasing function of the dimension p of Σ.

Following the results in van Emden (1971, p. 61) and Ljung and Rissanen (1978, p. 1421), and filling the gap in Maklad and Nichols (1980, p. 82), the proof and the properties of $C_1(\Sigma)$ in (25) are shown in detail by Bozdogan (1990, 1998a, 1998b).

The contribution of the complexity of the model covariance structure is that it provides a numerical measure to assess *parameter redundancy* and *stability* uniquely all in one measure. When the parameters are stable, this implies that the covariance matrix should be approximately a diagonal matrix. This concept of a stable parameter is equivalent to the simplicity of a model covariance structure defined in Bozdogan (1988a, 1988b). Indeed, $C_1(\Sigma)$ penalizes the scaling of the ellipsoidal dispersion, and the importance of circular distribution has been taken into account. It is because of these reasons that we use $C_1(\Sigma)$ without using any transformations of Σ and that we do not discard the use of $C_0(\Sigma)$.

Let $\lambda_1, \lambda_2, ..., \lambda_p$ be the eigenvalues of Σ, then $tr(\Sigma)/p = \bar{\lambda}_a = 1/p \sum_{j=1}^p \lambda_j$ is the arithmetic mean of the eigenvalues of Σ, and $|\Sigma|^{1/p} = \bar{\lambda}_g = (\prod_{j=1}^p \lambda_j)^{1/p}$ is the geometric mean of the eigenvalues of Σ. Then the complexity of Σ can be written as

$$C_1(\Sigma) = \frac{p}{2}\log(\bar{\lambda}_a/\bar{\lambda}_g). \tag{26}$$

Hence, we interpret the complexity as the log ratio between the arithmetic mean and the geometric mean of the eigenvalues of Σ. It measures how unequal the eigenvalues of Σ are, and it incorporates the two simplest scalar measures of multivariate scatter, namely the *trace* and the *determinant* into one single function. Indeed, Mustonen (1997) in a recent paper studies the fact that the *trace* (*sum of variances*)

and the *determinant* of the covariance matrix Σ (*generalized variance*) alone do not meet certain essential requirements of variability in the multivariate normal distribution.

In general, large values of complexity indicate a high interaction between the variables, and a low degree of complexity represents less interaction between the variables. The minimum of $C_1(\Sigma)$ corresponds to the *least complex* structure. In other words, $C_1(\Sigma) \to 0$ as $\Sigma \to I$, the identity matrix. This establishes a plausible relation between information-theoretic complexity and computational effort. Furthermore, what this means is that the identity matrix is the least complex matrix. To put it in statistical terms, orthogonal designs, or linear models with no colinearity, are the least complex, or most informative, and the identity matrix is the only matrix for which the complexity vanishes. Otherwise, $C_1(\Sigma) > 0$, necessarily.

Geometrically, $C_1(\Sigma)$ preserves all inner products, angles, and lengths under orthogonal transformations of Σ. An orthogonal transformation T indeed exists which corresponds to a sequence of plane rotations of the coordinate axes to equalize the variances. This can be achieved using *Jacobi's iterative method* or *the Gauss–Seidel method* (see, Graham, 1987).

We note that the system correlation matrix can also be used to describe complexity. If we wish to show the interdependencies (i.e., *correlations*) among the parameter estimates, then we can transform the covariances to correlation matrices and describe yet another useful measure of complexity. Let $\mathbf{R}$ be the correlation matrix obtained from Σ by the relationship $\mathbf{R} = \Lambda_\sigma \Sigma \Lambda_\sigma$, where $\Lambda_\sigma = diag(1/\sigma_1, ..., 1/\sigma_p)$ is a diagonal matrix whose diagonal elements equals $1/\sigma_j$, $j = 1, 2, ..., p$. From (25), we have $C_1(\mathbf{R}) = -1/2 \log |\mathbf{R}| \equiv C_0(\mathbf{R})$. The diagonal operation of a covariance matrix Σ always reduces the complexity of Σ, and $C_1(\mathbf{R}) \equiv C_0(\mathbf{R})$ takes into account the interdependencies (correlations) among the variables. For simplicity, the C_0 measure based on the correlation matrix $\mathbf{R}$ will be denoted by C_R and $C_0(\mathbf{R})$ is written as $C_R(\Sigma)$ for notational convenience, since $\mathbf{R}$ is obtained from Σ. Obviously, C_R is invariant with respect to scaling and orthonormal transformations and subsequently can be used as a complexity measure to evaluate the interdependencies among parameter estimates. Note that if $|\mathbf{R}| = 1$, then $I(x_1, x_2, ..., x_p) = 0$ which implies the mutual independence of the variables $x_1, x_2, ..., x_p$. If the variables are not mutually independent, then $0 < |\mathbf{R}| < 1$ and $I(x_1, x_2, ..., x_p) > 0$. In this sense $I(\mathbf{x})$ in (20) or (21) can also be viewed as a measure of dimensionality of model manifolds.

Next, we develop the information complexity ICOMP(IFIM) approach to model evaluation based on the maximal covariance complexity $C_1(\bullet)$, and $C_R(\bullet)$.

ICOMP: A NEW INFORMATION MEASURE OF COMPLEXITY FOR MODEL SELECTION

In this section, we introduce a new model-selection criterion called ICOMP(IFIM) to measure the fit between multivariate normal linear *and/or* nonlinear structural models and observed data as an example of the application of the covariance complexity measure defined in the previous section. ICOMP(IFIM) resembles a penalized likelihood method similar to AIC and AIC-*type criteria*, except that the penalty

depends on the curvature of the log likelihood function via the scalar $C_1(\bullet)$ complexity value of the *estimated* IFIM.

ICOMP as an Approximation to the Sum of Two Kullback–Leibler Distances

DEFINITION 7. For a multivariate normal linear or nonlinear structural model we define the general form of ICOMP(IFIM) as

$$\text{ICOMP(IFIM)} = -2 \log L(\hat{\theta}) + 2C_1(\hat{\mathscr{F}}^{-1}(\hat{\theta})), \tag{27}$$

where C_1 denotes the maximal information complexity of $\hat{\mathscr{F}}^{-1}$, the estimated IFIM.

To show this, suppose we consider a general statistical model of the form given by

$$\mathbf{y} = m(\theta) + \varepsilon, \tag{28}$$

where $\mathbf{y} = (y_1, y_2, ..., y_n)$ is an $(n \times 1)$ random vector of response values in $\Re^n$; θ is a parameter vector in $\Re^k$; $m(\theta)$ is a systematic component of the model in $\Re^n$, which depends on the parameter vector θ, and its deterministic structure depends on the specific model considered, e.g., in the usual linear multiple regression model $m(\theta) = \mathbf{X}\theta$, where $\mathbf{X}$ is an $(n \times (k+1))$ matrix of nonstochastic or constant design or model matrix with k explanatory variables so that $rank(\mathbf{X}) = k + 1 = q$; and ε is an $(n \times 1)$ random error vector with

$$\mathscr{E}(\varepsilon) = \mathbf{0}, \qquad \text{and} \qquad \mathscr{E}(\varepsilon\varepsilon') = \Sigma_\varepsilon. \tag{29}$$

We denote θ^* to be vector of parameters of the operating true model and θ to be any other value of the vector of parameters. Let $f(\mathbf{y}; \theta)$ denote the joint density function of $\mathbf{y}$ given θ. Let $I(\theta^*; \theta)$ denote the *KL distance* between the densities $f(\mathbf{y}; \theta^*)$ and $f(\mathbf{y}; \theta)$. Then, since y_i are independent, $i = 1, 2, ..., n$, we have

$$\begin{aligned} I(\theta^*; \theta) &= \int_{\Re^n} f(\mathbf{y}; \theta^*) \log\left[\frac{f(\mathbf{y}; \theta^*)}{f(\mathbf{y}; \theta)}\right] d\mathbf{y} \\ &= \sum_{i=1}^{n} \int f_i(y_i; \theta^*) \log[f_i(y_i; \theta^*)]\, dy_i - \sum_{i=1}^{n} \int f_i(y_i; \theta^*) \log[f_i(y_i; \theta)]\, dy_i, \end{aligned} \tag{30}$$

where f_i, $i = 1, 2, ..., n$ are the marginal densities of the y_i.

Note that the first term in (30) is the usual negative entropy $H(\theta^*; \theta^*) \equiv H(\theta^*)$ which is constant for a given $f_i(y_i; \theta^*)$. The second term is equal to

$$-\sum_{i=1}^{n} \mathscr{E}[\log f_i(y_i; \theta)], \tag{31}$$

which can be unbiasedly estimated by

$$-\sum_{i=1}^{n} \log f_i(y_i; \theta) = -\log L(\theta \mid y_i), \tag{32}$$

where $\log L(\theta \mid y_i)$ is the log likelihood function of the observations evaluated at θ. Given a model M where the parameter vector is restricted, a maximum likelihood estimator $\hat{\theta}_M$ can be obtained for θ and the quantity

$$-2 \sum_{i=1}^{n} \log f_i(y_i; \hat{\theta}_M) = -2 \log L(\hat{\theta}_M)$$

evaluated. This will give us the estimation of the first KL *distance* which is reminiscent to the derivation of AIC. On the other hand, a model M gives rise to an asymptotic covariance matrix $Cov(\hat{\theta}_M) = \Sigma(\hat{\theta}_M)$ for the MLE $\hat{\theta}_M$. That is,

$$\hat{\theta}_M \sim N(\theta^*, \Sigma(\hat{\theta}_M) \equiv \mathscr{F}^{-1}(\hat{\theta}_M)). \tag{33}$$

Now invoking the $C_1(\bullet)$ complexity on $\Sigma(\hat{\theta}_M)$ from the previous section can be seen as the KL *distance* between the joint density and the product of marginal densities for a normal random vector with covariance matrix $\Sigma(\hat{\theta}_M)$ via (21), maximized over all orthonormal transformations of that normal random vector (see Bozdogan, 1990). Hence, using the estimated covariance matrix, we define ICOMP as the sum of two KL *distances* given by

$$\begin{aligned} \text{ICOMP(IFIM)} &= -2 \sum_{i=1}^{n} \log f_i(y_i; \hat{\theta}_M) + \mathbf{2} C_1(\hat{\Sigma}(\hat{\theta}_M)) \\ &= -2 \log L(\hat{\theta}_M) + \mathbf{2} C_1(\hat{\mathscr{F}}^{-1}(\hat{\theta}_M)). \end{aligned} \tag{34}$$

The first component of ICOMP(IFIM) in (34) measures the lack of fit of the model, and the second component measures the complexity of the estimated IFIM, which gives a scalar measure of the celebrated *Cramér–Rao lower bound matrix* which takes into account the accuracy of the estimated parameters and implicitly adjusts for the number of free parameters included in the model.

This approach has several rather attractive features. If $\mathscr{F}_{jj}^{-1}(\theta_K)$ is the jth diagonal element of the IFIM, from Chernoff (1956), we know that $\mathscr{F}_{jj}^{-1}(\bullet)$ represents the variance of the asymptotic distribution of $\sqrt{n}\,(\hat{\theta}_j - \theta_j)$, for $j = 1, ..., K$. Considering a subset of the K parameters of size k, we have that

$$\mathscr{F}_{jj}^{-1}(\theta_K) \geqslant \mathscr{F}_{jj}^{-1}(\theta_k). \tag{35}$$

Behboodian (1964) explains that the inequality (35) means that the variance of the asymptotic distribution of $\sqrt{n}\,(\hat{\theta}_j - \theta_j)$ can only increase as the number of unknown parameters is increased. The use of the $C_1(\hat{\mathscr{F}}^{-1}(\hat{\theta}))$ in the information-theoretic model evaluation criteria takes into account the fact that as we increase the number of free parameters in a model, the accuracy of the parameter estimates decreases. As

preferred according to the principle of parsimony, ICOMP(IFIM) chooses simpler models that provide more accurate and efficient parameter estimates over more complex, overspecified models. Further, we note that in (34), the *trace* of IFIM in the complexity measure involves only the diagonal elements analogous to *variances* while the *determinant* involves also the off-diagonal elements analogous to *covariances.* Therefore, ICOMP(IFIM) contrasts the *trace* and the *determinant* of IFIM, and this amounts to a comparison of the *geometric* and *arithmetic means* of the *eigenvalues* of IFIM as shown in (26). The *greatest simplicity*, that is *zero complexity*, is achieved when IFIM is proportional to the identity matrix, implying that the *parameters are orthogonal* and can be estimated with equal precision. In this sense, *parameter orthogonality*, several forms of *parameter redundancy*, and *parameter stability* are all taken into account.

We note that ICOMP(IFIM) in (34) penalizes the *bad scaling* of the parameters. It is important to note that good conditioning of the information matrix needs a simple structure, but the latter does not necessarily imply the former. For example, consider an information matrix which is diagonal with some diagonal elements close to zero. In this case, the corresponding correlation matrix is an identity matrix, which is the simplest. But, the information matrix is poorly conditioned. Therefore, the analysis based on the correlation matrix ignores an important characteristic, namely, the ratios of the diagonal elements in the information matrix, or the *scale* of these components.

If scale invariance is an issue in model selection enterprise, then one can use the *correlational form of* IFIM, that is, $\mathscr{F}_R^{-1}(\hat{\theta}) = D_{\mathscr{F}^{-1}}^{-1/2} \mathscr{F}^{-1} D_{\mathscr{F}^{-1}}^{-1/2}$, and get

$$\text{ICOMP(IFIM)}_R = -2 \log L(\hat{\theta}) + 2C_1(\hat{\mathscr{F}}_R^{-1}(\hat{\theta})). \tag{36}$$

Parameter transformation can reduce the complexity measure based on the correlation structure, but it can increase the complexity measure based on the maximal complexity. This occurs because the reduction in the correlation does not imply the reduction of scaling effect. Indeed, the reduction in the correlation may even make scaling worse as we described above. In this sense, ICOMP(IFIM) may be better than ICOMP(IFIM)_R, especially in nonlinear models, since it considers both of these effects in one criterion. For more on the above, see, e.g., Chen (1996), Chen and Bozdogan (1999), and Bozdogan(1998a, 1998b).

With ICOMP(IFIM), complexity is viewed not as the number of parameters in the model, but as the *degree of interdependence* (i.e., the *correlational structure* among the parameter estimates). By defining complexity in this way, ICOMP(IFIM) provides a more judicious penalty term than AIC, Rissanen's (1978, 1986) MDL, Schwarz's (1978) SBC (*or* BIC), and Bozdogan's (1987a) CAIC. The lack of parsimony is automatically adjusted by $C_1(\hat{\mathscr{F}}^{-1}(\hat{\theta}))$ across the competing alternative portfolio of models as the parameter spaces of these models are constrained in the model selection process.

We give the *relative reduction of complexity* (*RRC*) in terms of the *estimated* IFIM as

$$\text{RRC} = \frac{C_1(\hat{\mathscr{F}}^{-1}(\hat{\theta})) - C_1(\hat{\mathscr{F}}_R^{-1}(\hat{\theta}))}{C_1(\hat{\mathscr{F}}^{-1}(\hat{\theta}))}. \tag{37}$$

The *percent relative reduction of complexity* is then given by

$$\text{PRRC} = \frac{C_1(\hat{\mathcal{F}}^{-1}(\hat{\theta})) - C_1(\hat{\mathcal{F}}_R^{-1}(\hat{\theta}))}{C_1(\hat{\mathcal{F}}^{-1}(\hat{\theta}))} \times 100\,\%. \tag{38}$$

The RRC or the PRRC, gives us a yard stick of how to determine which models are indeed the best fitting model(s) across a portfolio of alternative models. Indeed, the interpretation of RRC or PRRC is that they both measure heteroscedastic complexity plus a correlational complexity of the model.

As discussed in Morgera (1985, p. 612) all the complexity of a covariance matrix is manifasted in the off-diagonal elements only; whereas, the matrices **R** which are not in Toeplitz form are not variance equalized.

There are other formulations of ICOMP which are based on the covariance matrix properties of the parameter estimates of a model starting from their finite sampling distributions. These versions of ICOMP are useful in linear models. For example, for a multivariate normal linear or nonlinear structural model, under the assumption that the estimation of the parameters in the expectation is independent of the estimation of the covariance structure of the errors, the model complexity is defined as

$$C_1(\hat{\Sigma}_{model}) = C_1(\hat{\Sigma}_\theta) + C_1(\hat{\Sigma}_\varepsilon). \tag{39}$$

Then ICOMP is defined as

$$\text{ICOMP} = -2\log L(\hat{\theta}) + 2[\,C_1(\hat{\Sigma}_\theta) + C_1(\hat{\Sigma}_e)\,]. \tag{40}$$

The first component of ICOMP in (40) measures the lack of fit, the second component measures the complexity of the covariance matrix of the parameter estimates of a model, and the third component measures the complexity of the covariance matrix of the model residuals. We note that if the random error ε is assumed to be normally distributed and spherical, i.e., if $\varepsilon \sim N(0, \sigma^2\mathbf{I})$, then the third component of ICOMP in (40) will be zero in the usual multiple regression models, since the covariance matrix of the projection matrix in local coordinates is $\sigma^2\mathbf{I}_{n-q}$, and we also note that the complexity of $\sigma^2\mathbf{I}_{n-q}$ is zero. In other words $C_1(\sigma^2\mathbf{I}_{n-q}) = 0$. This state of affairs is due to the dubious assumption that the random errors have $\sigma^2\mathbf{I}$ as their covariance matrix.

To take the scale invariancy into account and to show the *interdependencies* (i.e., *correlations*) among the parameter estimates, we can further define the correlational form of ICOMP in (40) and get

$$\text{ICOMP}_R = -2\log L(\hat{\theta}) + 2[\,C_R(\hat{R}_\theta) + C_R(\hat{R}_e)\,]. \tag{41}$$

For more details, we refer the readers to Bozdogan (1998a, 1998b).

Next, we give our numerical examples to illustrate some of the points discussed in this paper and conclude with a discussion of the comparison of information criteria.

NUMERICAL EXAMPLES

In this section, we give several numerical examples to demonstrate the practical utility of the proposed model selection criteria.

EXAMPLE 1. *Improving Parameter Stability*. Parameter stability plays an important role in statistical modeling. The concept of parameter stability was proposed by Ross (1970). The basic idea is that the parameters characterizing a model should be chosen so that they are affected little by changes in the remaining parameters. As we discussed before, parameter stability implies that the covariance matrix of the parameters or the parameter estimates should approximate a diagonal matrix.

One important technique for improving parameter stability is parameter transformation. To illustrate the impact of parameter transformation on parameter stability, we consider the example given in Chen (1996, p. 28) and Chen and Bozdogan (1998) which was originally motivated by Ross (1970) and later was reconsidered by Seber and Wild (1989) in a simulation study.

The proposed simple nonlinear regression model is

$$M_1: y = \beta(1 - e^{\gamma x}) + \varepsilon \tag{42}$$

with the simulated data given in Table 1.

The model is fitted with estimated error variance $\hat{\sigma}^2 = 0.002146$. The contour of the 95 and 99% confidence regions of the parameter estimates of model M_1 is shown in Fig. 1. We note that this contour is highly curved, which implies a high intercorrelation between the estimates of β and γ.

We consider two sets of parameter transformations of model M_1.

Transformation 1: $\theta_1 = \beta\gamma$, $\theta_2 = \gamma$. The corresponding model is

$$M_2: y = \frac{\theta_1}{\theta_2}(1 - e^{\theta_2 x}) + \varepsilon. \tag{43}$$

The corresponding contour of the 95 and 99% confidence regions of the parameter estimates of model M_2 is shown in Fig. 2.

The parameter transformation has turned a highly curved contour into one close to an ellipsoid. But, the axes of this contour are not parallel to the coordinate axes, so θ_1 and θ_2 are still highly correlated to each other.

TABLE 1

Simulated Data

y	0.0638	0.1870	0.2495	0.3207	0.3356	0.5040	0.5030	0.6421	0.6412	0.5678
x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

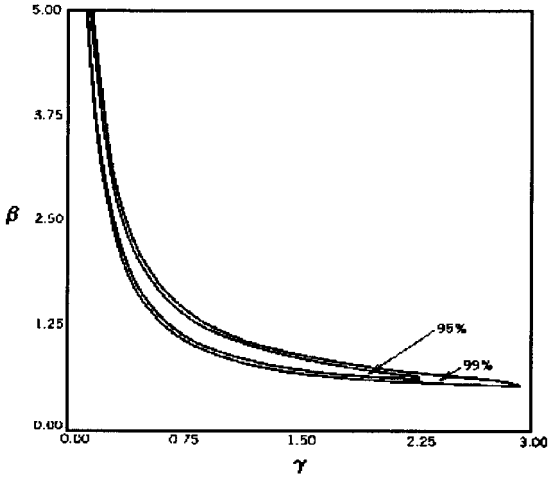

FIG. 1. Exact confidence regions of model M_1.

Transformation 2. $\alpha_1 = \beta\gamma$, $\alpha_2 = \beta\gamma - 0.35\gamma$. The corresponding model becomes

$$M_3: y = \frac{0.35\alpha_1}{\alpha_1 - \alpha_2}\left(1 - \exp\left(-\frac{(\alpha_1 - \alpha_2)x}{0.35}\right)\right) + \varepsilon. \tag{44}$$

The corresponding contour of the 95 and 99% confidence regions of the parameter estimates of model M_3 is shown in Fig. 3.

This set of parameter transformations has turned the curved contour into a contour which is close to an ellipsoid with axes nearly parallel to the coordinate axes of the new parameters.

We note that all three models fit data equally well in the sense that they have the same estimated error variance $\hat{\sigma}^2 = 0.002146$, or say, they reach the same likelihood function value. But we further note that the qualities of these models are significantly different due to the shape of the contours. To see the evaluation of parameter stabilities, complexity measures for all three models are computed and summarized in Table 2.

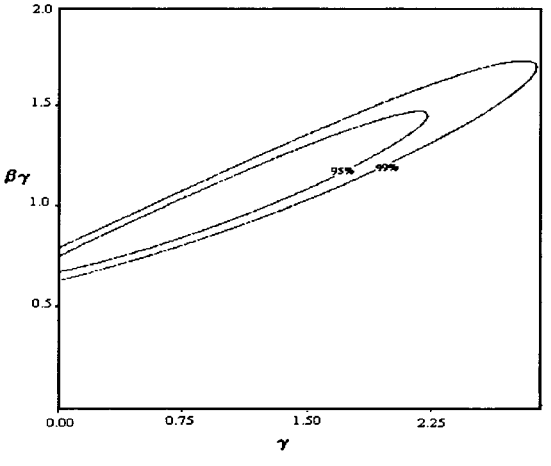

FIG. 2. Exact confidence regions of model M_2.

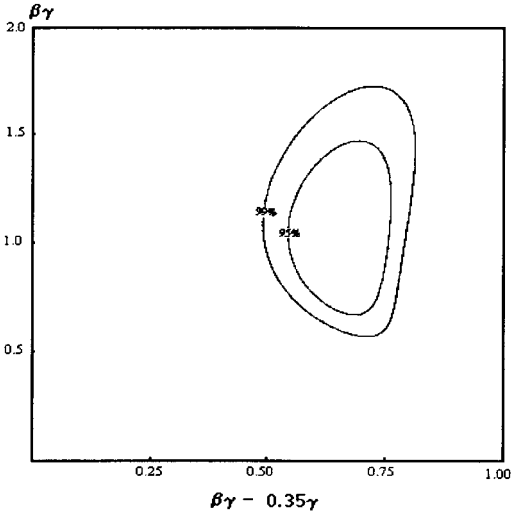

FIG. 3. Exact confidence regions of model M_3.

Examining the results in Table 2, we see that among the three models, model M_3 has the simplest covariance structure. Moreover, since all three models have an equal lack of fit, ICOMP criteria will simply select the simplest model, which is model M_3.

This example demonstrates the importance of parameter transformation in reducing model complexity. Although the complexity can be seen visually through the shape of the contours, such contours will not be available to the researcher when the dimension of the parameter space is larger than two. In this case, the C_1 and C_R measures provide useful tools to evaluate complexity. Further, in this example we also note that the original parameters and the transformed parameters have the same dimension. In general, the transformed parameters may not necessarily have the same dimension as the original parameters. In this case, it makes sense to consider the situation where the transformed parameters have a higher dimension as compared to the original parameters, which significantly reduces the lack of fit as the dimension of the parameters increases.

EXAMPLE 2. *Subset Selection of Variables in Multivariate Regression.* In many applications in behavioral and social sciences, econometric modeling, environmental sciences, and many other fields, it is often the case that several dependent variables are simultaneously considered as one target with a set of independent variables. Often

TABLE 2

Complexity of the Transformed Models

Model	$C_1(\hat{\Sigma}_\theta)$	$C_R(\hat{\Sigma}_\theta)$
M_1	7.7528	2.1917
M_2	7.2076	1.3702
M_3	4.9268	0.00036

predicting all the target dependent variables simultaneously from the set of independent variables are desired to determine the best predictors.

For space considerations, here we consider a small multivariate regression data set from Finn (1974, p. 67) to select the best predictors of creativity and achievements of $n = 15$ freshmen class at a large midwestern university. The response (or dependent) variables are: $y_1 =$ grade average for required courses taken, $y_2 =$ grade average for elective courses taken; and the independent variables are: $x_1 =$ high school general knowledge test, $x_2 =$ IQ score from previous year, and $x_3 =$ educational motivation score from previous year. We carry out a subset selection of variables using the multivariate regression model

$$\mathbf{Y}_{(n\times p)} = \mathbf{X}_{(n\times q)}\mathbf{B}_{(q\times p)} + \mathbf{E}_{(n\times p)} \tag{45}$$

with a no-constant term. In (45), $n = 15$, $p = 2$, and $q = k = 3$, for the *no-constant model*. We derive and score information theoretic criteria under the multivariate normal assumption for the model in (45) which are given as follows.

$$\text{AIC(Multivar Re } g) = np\log(2\pi) + n\log|\hat{\Sigma}| + np + 2\left[pq + \frac{p(p+1)}{2}\right]; \tag{46}$$

$$\text{ICOMP(IFIM)}_{\textit{Multivar Re } g} = np\log(2\pi) + n\log|\hat{\Sigma}| + np + 2C_1(\hat{\mathscr{F}}^{-1}(\hat{\Theta})), \tag{47}$$

where the *estimated* IFIM, following Magnus and Neudecker (1988), for the multivariate regression model is given by

$$\hat{\mathscr{F}}^{-1}(\hat{\Theta}) = \begin{bmatrix} \hat{\Sigma}\otimes(\mathbf{X}'\mathbf{X})^{-1} & 0 \\ 0 & \frac{2}{n}D_p^{+}(\hat{\Sigma}\otimes\hat{\Sigma})D_p^{+\prime} \end{bmatrix}. \tag{48}$$

In (48), D_p^{+} is the Moore–Penrose inverse of the duplication matrix D_p. Duplication matrix D_p is a unique $p^2 \times 1/2(p(p+1))$ matrix which transforms $v(\hat{\Sigma}) \equiv vech(\hat{\Sigma})$ into $vec(\hat{\Sigma})$.

The complexity measure $C_1(\hat{\mathscr{F}}^{-1}(\hat{\Theta}))$ then becomes:

$$C_1(\hat{\mathscr{F}}^{-1}(\hat{\Theta})) = \frac{p(p+q)}{2}\log\left[\frac{tr(\hat{\Sigma})\,tr(\mathbf{X}'\mathbf{X})^{-1} + \frac{1}{2n}\left[tr(\hat{\Sigma}^2) + (tr\,\hat{\Sigma})^2 + 2\sum_j \hat{\sigma}_{jj}^2\right]}{p(p+q)}\right]$$
$$-\frac{1}{2}(p+q+1)\log|\hat{\Sigma}| - \frac{p}{2}\log|(\mathbf{X}'\mathbf{X})^{-1}| - \frac{p}{2}\log(2). \tag{49}$$

Also,

$$\text{ICOMP(Multivar Re } g) = np\log(2\pi) + n\log|\hat{\Sigma}| + np$$
$$+2[(n+q)\,C_1(\hat{\Sigma}) + pC_1((\mathbf{X}'\mathbf{X})^{-1})]. \tag{50}$$

In the above equations, $\hat{\Sigma}$ is the estimated error covariance matrix of the model in (45). Note that the complexity measure C_1 avoids the full numerical construction of the *estimated* IFIM in (48) and in general. This we like, since for large complex models the dimension of the *estimated* IFIM will be quite large. Complexity C_1 (or C_R) helps reduce the cost of computing large matrices, since it produces the scalar measure of the matrix. For the full construction of IFIM of any dimension, see Bozdogan (1990, 1994d, 1998a, 1998b), and Magnus and Neudecker (1988). We also compute ICOMP(IFIM)$_R$ from (36), and the *percent relative reduction in complexity* given in (38).

Our results in carrying out the subset selection of variables to determine the best predictors of performance of 15 college freshmen using a multivariate regression model with a no-constant term are summarized in Table 3.

We observe that all the criteria choose the subset $\{x_1, x_2\}$ to be the best set of predictors at 2-*level*. Note that AIC achieves its global minima at this subset level. So, $x_1 =$ high school general knowledge test and $x_2 =$ IQ score from previous year are the best predictors of performance at the 2-*level* subset. On the other hand, ICOMP(IFIM)$_R$ and ICOMP choose the subset $\{x_1\}$ as the best singleton subset at 1-*level*. It is important to note that the highest percent of relative reduction of the complexity occurs also at the subset $\{x_1\}$, which gives us an added yardstick for the best fitting model chosen.

Our last example, is a Monte Carlo simulation result to illustrate the performance of AIC and ICOMP *class criteria* along with SBC/MDL.

EXAMPLE 3. Choosing the Lag Order in Vector Autoregressive Models. Multivariate time series models known also as *vector autoregressive* (VAR) model provide a convenient method for forecasting time series data. A vector autoregression of order k, denoted VAR(k), represents the unrestricted reduced form of a dynamic structural model which is of the same form as the multivariate regression model given in (41), with the exception that all the predictors are generated from the left-hand side within the VAR model itself. In this example, our goal is to illustrate the performance of the information-based criteria in choosing the correct

TABLE 3

k	Subset of variables	m^c	ICOMP(IFIM)	ICOMP(IFIM)$_R$	ICOMP	AIC	PRRC
3	$\{x_1, x_2, x_3\}$	9	71.0681	61.4503	76.6753	64.6608	39.41%
2	$\{x_1, x_2\}$	7	**70.9728**[a]	**60.7443**[a]	**68.6383**[a]	**61.6862**[a, b]	**43.92%**
2	$\{x_1, x_3\}$	7	**70.5289**[a]	65.3564	76.2266	66.1531	28.15%
2	$\{x_2, x_3\}$	7	76.2974	69.2399	80.8667	70.5329	35.71%
1	$\{x_1\}$	5	71.6129	**56.5659**[a, b]	**62.5323**[a, b]	**64.2965**[a]	**85.89%**
1	$\{x_2\}$	5	78.7327	62.0653	69.9769	68.9259	84.15%
1	$\{x_3\}$	5	**68.0575**[a, b]	59.9816	65.9816	66.6155	76.60%

[a] Best subset variables at each level.

[b] Global minima of the criteria.

[c] Number of parameters.

TABLE 4

Criteria/order	0	1	**2***	3	4	5	6
AIC	0.000	0.006	**0.870**	0.084	0.032	0.006	0.002
SBC/MDL	0.000	0.274	**0.724**	0.002	0.000	0.000	0.000
ICOMP(IFIM)	0.000	0.002	**0.908**	0.046	0.016	0.004	0.006

order of the $VAR(k)$ model. The *true model* is from Lutkepohl (1991, p. 119) which is a stationary $VAR(k^* = 2)$ with $p = 2$ variables given by

$$y_t = \begin{bmatrix} 0.500 & 0.100 \\ 0.400 & 0.500 \end{bmatrix} y_{t-1} + \begin{bmatrix} 0.000 & 0.000 \\ 0.250 & 0.000 \end{bmatrix} y_{t-2} + \varepsilon, \tag{51}$$

where

$$\varepsilon \sim N_{p=2}\left(\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 0.90 & 0.30 \\ 0.30 & 0.4 \end{bmatrix}\right). \tag{52}$$

Note that the covariance matrix of the random error in (52) is not a diagonal matrix indicating the existence of the correlations among the components of the random error matrix. This will make it difficult for the model selection criteria to choose the *true* $VAR(k^* = 2)$ model almost surely.

We simulated a total of $n = 200$ observations from the above *true* $VAR(k^* = 2)$ *model* and replicated the Monte Carlo experiment 500 times. Then, we fitted $VAR(k)$ *models* with varying order $k = 0, 1, 2^*, ..., 6$, to the generated data sets, and scored the information criteria. Our results from this experiment are summarized in Table 4, which gives the relative frequency of the estimated *true stationary* $VAR(k^* = 2)$ *model* chosen by each of the model selection criteria.

Looking at Table 4, we see that AIC picks the true order $k = 2^*$, 87% of the time, and there is some overfitting which is not surprising for AIC, and almost no underfitting. SBC/MDL criteria pick the true order 72.4% of the time, and there is 27.4% underfitting by these two criteria with almost no overfitting. On the other hand ICOMP(IFIM) chooses the correct order of the *VAR* model 90.8% of the time with almost no underfitting, and with very little overfitting. Indeed, ICOMP(IFIM) performs better than AIC, and SBC/MDL type criteria in this experiment and controls the overfitting and underfitting risks judiciously.

For more on VAR models, see Bearse and Bozdogan (1998) and Bozdogan and Bearse (1998).

CONCLUSION: COMPARISON OF INFORMATION CRITERIA

In this paper, we introduced several information criteria, namely, AIC and its *variants*, and a new set of criteria, ICOMP(IFIM), ICOMP(IFIM)$_R$, ICOMP, and ICOMP$_R$, as alternative procedures. It is important to recognize the differences among these criteria.

Among these criteria, we briefly studied the underlying basic idea of the AIC procedure and presented its derivation as a bias correcting criterion. As it is well known, AIC simply defines the model complexity in terms of the number of free parameters, so that it is the easiest to apply. Although AIC is invariant under parameter transformations, it does not have the virtue of detecting the problem caused by the curvature of a model, especially in univariate and multivariate nonlinear models. Also as pointed out by many researchers, AIC often overfits the model which has been known to Akaike.

In the second part of this paper, we introduced a new entropic or information-theoretic measure complexity ICOMP criterion in several forms. Among these, ICOMP(IFIM) and ICOMP are based on the covariance matrices which have nice features. Besides the shape of the contour of the parameter space of a model, both ICOMPs based on the covariance matrices also penalize the scaling of the ellipsoidal dispersion of a model. For example, when an ellipsoid is parallel to the coordinates and has a long and narrow shape, it can indicate that we might have an *ill-conditioned Fisher information matrix*. With ICOMP we are able to detect the occurrence of such problems and find out the contour with a good shape among other candidate contours. Also with ICOMP we are able to measure how far we might be from a circular contour shape, which emphasizes the role of a good shape of a contour as we illustrated in Example 1.

The difference between the general approach and the finite sampling approach in developing ICOMP is worth noting. Besides penalizing the covariance complexity in estimating the parameter vector θ, the finite sampling approach penalizes the covariance complexity of the residual covariance matrix of the model. So this approach is useful to study cases where the random errors of the model might be correlated, which gives us the provision of unifying the situations where the model residuals might be both correlated and uncorrelated by including dependence. The general approach, that is, ICOMP(IFIM), penalizes covariance complexity of the estimated parameters of the entire model. It provides a *trade-off* between *lack of fit* and *a scalar measure of the accuracy of the parameter estimates.*

To make ICOMP(IFIM) and its variants be scale invariant, we then introduced their correlational forms, namely, ICOMP(IFIM)_R and ICOMP_R. These two correlational forms penalize the shape of the contour of the parameter space of a model by *trading-off* with the shape of the contour of the likelihood function of a model. As discussed by Ross (1970), a good shape of the contour should be close to an ellipsoid and the axes of the ellipsoid should be parallel to the coordinates (see Example 1).

Similar to AIC, we do not claim that ICOMP-*type criteria* are consistent. This is due to constant, or asymptotically constant, penalty functionals in model selection criteria. However, we emphasize the fact that the *constancy* in the complexity of ICOMP is quite different from that of AIC, since the definition of $C_1(\bullet)$ and $C_R(\bullet)$ explicitly takes into account the *interdependencies* (i.e., *correlations*) among the variables and both linearity and nonlinearity of the parameters of the model. Therefore, the *constancy* of ICOMP is not an integer; it varies. For example, in equivalent models, $2k$ and $klog(n)$ type penalty terms will be the same with no hope of distinguishing among the equivalent models. With the definition of $C_1(\bullet)$ and

$C_R(\bullet)$ it is now possible to distinguish among the equivalent models and control the risk of underfitting and overfitting phenomena judiciously. This is the major dilemma in model selection, not the issue of consistency. Nevertheless, consistency properties of ICOMP has been studied by Bozdogan and Haughton (1998) in the case of the usual multiple regression models, where the probabilities of underfitting and overfitting as the sample size n tends to infinity have been established. Through a large scale Monte Carlo *misspecification environment*, when the true model is not in the model set, the performance of ICOMP has been studied under different configurations of the experiment with varying sample sizes and the error variances. The results obtained show that ICOMP *class criteria* overwhelmingly agree most often with the KL decision which goes to the heart of the consistency arguments about information criteria not studied before, since most of the studies are based on the fact that the true model considered is in the model set.

Finally, we note that one of the advantages of ICOMP class criteria is that the rationale for combining goodness of fit terms with complexity measures does not rely on any regularity conditions. This means that, for example, ICOMP can help decide between one or two components in a mixture model (Bozdogan 1994d), a situation where regularity conditions fail notoriously. Furthermore, the difference between ICOMP class criteria and AIC, SBC/MDL, and CAIC is that with ICOMP we have the advantage of working with both biased as well as unbiased estimates of the parameters and measure the complexities of their covariances to study the *robustness properties* of different methods of parameter estimates. AIC and AIC-*type criteria* are based on MLE's, which often are biased and they do not fully take into account the concept of parameter redundancy, accuracy, and the parameter interdependencies in model fitting and selection process. Also, ICOMP class criteria legitimizes the role of the *Fisher information matrix* (FIM) as the natural metric on the parameter manifold of the model which remained academic for a long time. In the literature (see, e.g., Li, Lewandowsky, and DeBrunner, 1996, and others), a measure of a model's total sensitivity to all of its parameters is often defined in terms of the *trace of* FIM, and in some cases it is defined by the *determinant of* IFIM, called a generalized variance. Using such measures alone as performance measures has serious disadvantages to which one should pay attention (see Mustonen, 1997). On these and related problems, a joint work with Dr. In Jae Myung is currently underway and the results of this work will be published upon completion.

In conclusion, we note that our numerical results clearly demonstrate the excellent performance of ICOMP *class criteria* as compared to AIC and SBC/MDL to be used in model selection, prediction, and perturbation studies. We believe the set of potentially fruitful applications of information theoretic model selection criteria are vast in experimental and mathematical psychology, in psychometrics, social, and behavioral and economic sciences. We hope that future research will continue to explore these avenues.

ACKNOWLEDGMENTS

This research was partially supported by the Faculty Development Award at the University of Tennessee from January 1997 to August 1998. The author is grateful to many colleagues, including

Professors In Jae Myung and Malcolm Forster for inviting me to speak on the work presented in this paper at the *Symposium On Methods For Model Selection* at Indiana University, Bloomington, Indiana, August 3–4, 1997. Thanks are also due to Professor Richard M. Shiffrin for hosting such a wonderful symposium on a very important topic. Also, helpful comments of two anonymous referees are gratefully acknowledged.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Academiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, **16**, 3–14.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, **52**, 317–332.

Balasubramanian, V. (1996). A geometric formulation of Occam's razor for inference of parametric distributions. Unpublished paper available as preprint number adap-org/9601001 from http://xyz.lanl.gov/ and as Princeton University Physics Preprint PUPT-1588, January 1996.

Baxter, R. A. (1996). *Minimum message length inference*: *Theory and applications*. Unpublished doctoral dissertation, Department of Computer Science, Monash University, Clayton, Victoria, Australia.

Bearse, P. M., & Bozdogan, H. (1998). Subset selection in vector autoregressive (VAR) models using the genetic algorithm with informational complexity as the fitness function. *Systems Analysis*, *Modeling*, *Simulation (SAMS)*, **31**, 61–91.

Behboodian, J. (1964). *Information for estimating the parameters in mixures of exponential and normal distributions*. Unpublished doctoral dissertation, Department of Mathematics, Univerisity of Michigan, Ann Arbor.

Bhansali, R. J., and Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika*, **64**, 547–551.

Boltzmann, L. (1877). Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respective den Satzen über das Wärmegleichgewicht. *Wiener Berichte*, **76**, 373–435.

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

Bozdogan, H. (1987b). ICOMP: A new model-selection criterion. Preprint paper in Hans H. Bock (Ed.), *Classification and related methods of data analysis*. Amsterdam: Elsevier Science (North-Holland).

Bozdogan, H. (1988a). ICOMP: A new model-selection criterion. In Hans H. Bock (Ed.), *Classification and related methods of data analysis*, pp. 599–608. Amsterdam: Elsevier Science (North-Holland).

Bozdogan, H. (1988b). The theory and applications of information-theoretic measure of complexity (ICOMP) as a new model selection criterion. Unpublished research report, the Institute of Statistical Mathematics, Tokyo, Japan, and the Department of Mathematics, University of Virginia, Charlottesville, VA, March 1988.

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in statistics theory and methods*, **19**, 221–278.

Bozdogan, H. (1994a). Theory & methodology of time series analysis (Vol. 1). *Proceedings of the first US/Japan conference on the frontiers of statistical modeling*: *An informational approach*. Dordrecht: Kluwer Academic.

Bozdogan, H. (1994b). Multivariate statistical modeling (Vol. 2). *Proceedings of the first US/Japan conference on the frontiers of statistical modeling*: *An informational approach*, Dordrecht: Kluwer Academic.

Bozdogan, H. (1994c). Engineering & scientific applications of informational modeling (Vol. 3). *Proceedings of the first US/Japan conference on the frontiers of statistical modeling*: *An informational approach*. Dordrecht: Kluwer Academic.

Bozdogan, H. (1994d). Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In H. Bozdogan (Ed.), *Multivariate statistical modeling* (Vol. 2), pp. 69–113. *Proceedings of the first US/Japan conference on the frontiers of statistical modeling*: *An informational approach*. Dordrecht: Kluwer Academic.

Bozdogan, H. (1996). *A new informational complexity criterion for model selection*: *The general theory and its application*. Invited paper presented in the Session on Information Theoretic Models & Inference of the Institute for Operations Research & the Management Sciences INFORMS, Washington D.C., May 5–8, 1996.

Bozdogan, H. (1998a). *Statistical modeling and model evaluation*: *A new informational approach*, manuscript.

Bozdogan, H. (1998b). *Informational complexity and multivariate statistical modeling*, manuscript.

Bozdogan, H., & Bearse, P. M. (2000). Model selection using informational complexity with appplications to vector autoregressive (VAR) models. *Journal of Statistical Planning and Inference*, in preparation.

Bozdogan, H., & Haughton, D. M. A. (1998). Informational complexity criteria for regression models. *Computational Statistics and Data Analysis*, in press.

Chen, X. (1996). *Model selection in nonlinear regression analysis*. Unpublished doctoral dissertation, Management Science Program, The University of Tennessee, Knoxville, TN.

Chen, X., & Bozdogan, H. (1998). *Model selection in nonlinear regression analysis*: *A new informational complexity approach*. Working book.

Chernoff, H. (1956). Large sample theory: Parametric case. *Annals of Mathematical Statistics*, **27**, 1–22.

Cover, T. M., Gacs, P., & Gray, R. M. (1989). Kolmogorov's contributions to information theory and algorithmic complexity. *Annals of Probability*, **17**, 840–865.

Cox, D. R. (1984). Effective degrees of freedom and the likelihood ratio test. *Biometrika*, **71**, 487–493.

Forster, M. R. (1999). The new science of simplicity. In H. A. Keuzenkamp, M. McAleer, and A. Zellner (Eds.), *Simplicity*, *inference*, *and econometric modeling*. Cambridge, UK: Cambridge University Press.

Finn, J. D. (1974). *A general model for multivariate analysis*. New York: Holt Rinehart and Winston.

Graham, A. (1987). *Nonnegative matrices and applicable topics in linear algebra*. New York: Halsted Press.

Hannan, E. J. (1986). Remembrance of things past. In J. Gani (Ed.), *The craft of probabilistic modeling*. New York: Springer-Verlag.

Harris, C. J. (1978). An information theoretic approach to estimation. In M. J. Gregson (Ed.), *Recent theoretical developments in control* (pp. 563–590). London: Academic Press.

Hosking, J. R. M. (1980). Lagrange-multiplier tests of time-series models. *Journal of the Royal Statistics Society*, *(Series B)*, **42**, 170–181.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Kitagawa, G., & Gersch, W. (1996). *Smoothness priors analysis of time series*. New York: Springer-Verlag.

Kolmogorov, A. N. (1983). Combinatorial foundations of information theory and the calculus of probabilities. *Russian Math Surveys*, **38**, 29–40.

Konishi, S., & Kitagawa, G. (1996). Generalized information criteria in model-selection. *Biometrika*, **83**, 875–890.

Konishi, S. (1998). Statistical model evaluation and information criteria. In S. Ghosh (Ed.), *Multivariate*, *design and sampling*. New York: Marcel Dekker.

Kullback, S. (1968). *Information theory and statistics*. New York: Dover.

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

Li, S-C., Lewandowsky, S., & DeBrunner, V. E. (1996). Using parameter sensitivity and interdependence to predict model scope and falsifiability. *Journal of Experimental Psychology*, **125**, 360–369.

Ljung, L., & Rissanen, J. (1978). On canonical forms, parameter identifiability and the concept of complexity. In N. S. Rajbman (Ed.), *Identification and system parameter estimation* (pp. 1415–1426). Amsterdam: North-Holland.

Lutkepohl, H. (1993). *Introduction to multiple time series analysis*. New York: Springer-Verlag.

Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus*. New York: Wiley.

Maklad, M. S., & Nichols, T. (1980). A new approach to model structure discrimination. *IEEE Transactions on Systems, Man, and Cybernetics*, **10**, 78–84.

Morgera, S. D. (1985). Information theoretic covariance complexity and its relation to pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, **15**, 608–619.

Mustonen, S. (1997). A measure of total variability in multivariate normal distribution. *Computational Statistics and Data Analysis*, **23**, 321–334.

Rissanen, J. (1976). Minmax entropy estimation of models for vector processes. In R. K. Mehra and D. G. Lainiotis (Eds.), *System identification* (pp. 97–119). New York: Academic Press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, **14**, 465–471.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*, **14**, 1080–1100.

Rissanen, J. (1987). Stochastic complexity (with discussion). *Journal of Royal Statistical Society, (Series B)*, **49**, 223–239.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Teaneck, NJ: World Scientific.

Ross, G. J. S. (1970). The efficient use of function minimization in nonlinear maximum likelihood estimation. *Applied Statistics*, **19**, 205–221.

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

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, **52**, 333–343.

Seber, G. A. F., & Wild, C. J. (1989). Nonlinear regression. New York: Wiley.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technology Journal*, **27**, 379–423.

Shibata, R. (1989). Statistical aspects of model selection. In J. C. Willems (Ed.), *From data to modeling* (pp. 216–240). Berlin: Springer-Verlag.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods*, **A7**, 13–26.

Takeuchi, K. (1976). Distribution of information statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, **153**, 12–18. [In Japanese]

Theil, H., & Fiebig, D. G. (1984). *Exploiting continuity: maximum entropy estimation of continuous distributions*. Cambridge, MA: Ballinger.

van Emden, M. H. (1971). *An analysis of complexity* (Vol. 35). Amsterdam: Mathematical Centre Tracts.

Wallace, C. S., & Freeman, P. R. (1987). Estimation and inference by compact coding (with discussion). *Journal of Royal Statistical Society, (Series B)*, **49**, 240–265.

Wallace, C. S., & Dowe, D. L. (1993). MML estimation of the von Mises concentration parameter. Technical Report 93/193, Department of Computer Science, Monash University, Clayton 3168, Australia.

Watanabe, S. (1985). *Pattern recognition: human and mechanical*. New York: Wiley.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–26.

Woodroofe, M. (1982). On model selection and the arc sine law. *Annals of Statistics*, **10**, 1182–1194.

Received: November 26, 1997; revised: October 26, 1998