İstanbul Üniversitesi İşletme Fakültesi Dergisi
Istanbul University Journal of the School of Business Administration
Cilt/Vol:39, Sayı/No:2, 2010, 370-398
ISSN: 1303-1732 – www.ifdergisi.org © 2010

Invited Paper

A new class of information complexity (ICOMP) criteria with an application to customer profiling and segmentation

Hamparsum Bozdogan[1]
Department of Statistics, Operations and Management Science
The University of Tennessee, Knoxville, TN, U.S.A.

Abstract

This paper introduces several forms of a new class of information-theoretic measure of complexity criterion called *ICOMP* as a decision rule for model selection in statistical modeling to help provide new approaches relevant to statistical inference. The practical utility and the importance of *ICOMP* is illustrated by providing a real numerical example in data mining of mobile phone data for customer profiling and segmentation of mobile phone customers using a novel *multi-class support vector machine-recursive feature elimination* (*MSVM-RFE*) method. The approach proposed in this paper outperforms the classical discriminant analysis techniques over 32% in terms of misclassification error rate.

This is a remarkable achievement due to using *MSVM-RFE* hybridized with *ICOMP* that was not possible using other methods to classify the mobile phone customer data base as a new micro-marketing analytics. This should capture the attention of the mobile phone industry for more refined analysis of their data bases for customer management and retention.

Keywords: *ICOMP class of criteria, covariance complexity, estimated inverse-Fisher information matrix (FIM), model selection, multi-class support vector machine-recursive feature elimination (MSVM-RFE), customer profiling and segmentation.*

Bilgi karmaşıklığı (ICOMP) kriterinin yeni bir sınıfı ile müşteri profili oluşturma ve segmentasyonu uygulaması

Bu çalışma, ICOMP olarak adlandırılan bilgi karmaşıklığı kriterinin yeni bir sınıfının tanıtımını amaçlamaktadır. Bu kriter, istatistiksel modellemede yeni yaklaşımlara yardım sağlamaktadır ve en iyi modelin seçilmesinde bir karar kuralı olarak kullanılır. ICOMP'un önemi ve kullanımı, veri madenciliğinde yeni bir yöntem olan "çok sınıflı destek vektör makineleri"ni kullanarak (MSVM-RFE), müşteri profili oluşturma ve segmantasyonu uygulamasında örnek verilerek gösterilmiştir.

Bu çalışmada önerilen yeni modelleme, cep telefonu kullanan müşterilerin sınıflandırılmasında, klasik diskriminant analizine göre elde edilen yanlış sınıflandırma oranının %32'sinden daha iyi bir performans göstermiştir.

Bu sonuçlar, yeni bir mikro-pazarlama analiz yöntemi olarak kullanılabilir. Ayrıca bu sonuçlar veri tabanlarını daha iyi analizler yaparak sınıflandırmada daha çok müşteri kazanmak isteyen veya ellerindeki müşterileri kaybetmek istemeyen cep telefonu piyasasının dikkatini çekebilir.

[1]*bozdogan@utk.edu (H. Bozdoğan)*

***Anahtar Sözcükler**: Yeni ICOMP sınıfı kriterler, kovaryans karmaşıklığı, tahminlenmiş Fisher bilgi matrisi (FIM) tersi, model seçme, çok sınıflı destek vektör makineleri – yinelemeli özellikli eleme, müşteri profili ve segmantasyonu.*

1. Introduction and Purpose

In this paper, we shall introduce several forms of a new class information-theoretic measure of complexity criterion called *ICOMP* of Bozdogan [1-6] as a decision rule for model selection in statistical modeling to help provide new approaches relevant to statistical inference. In *ICOMP*, *I* is for *information* and *COMP* for *complexity* to distinguish it from other non-information theoretic complexity measures.

In general statistical modeling and model evaluation problems, the concept of model complexity plays an important role. At the philosophical level, complexity involves notions such as connectivity patterns, and the interactions of model components. Without a measure of *overall* model complexity, prediction of model behavior and assessing model quality is difficult. This requires detailed statistical analysis and computation to choose the best fitting model among a portfolio of competing models for a given finite sample.

The development of *ICOMP* has been motivated in part by Akaike's [7] classic *information criterion* (*AIC*) given by

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

where $L(\hat{\theta}_k)$ is the maximized likelihood function, $\hat{\theta}_k$ is the maximum likelihood estimate of the parameter vector θ_k under the model M_k, and $m(k)$ is the number of independent parameters estimated when M_k is the model, 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 [8].

For a general multivariate linear or nonlinear model defined by

Statistical Model=Signal+Noise,

ICOMP is designed to estimate a loss function

$$\left.\begin{array}{l}\left.\begin{array}{l} Lack\,of\ \ fit \\ Loss = +Lack\,of\ Parsimony \end{array}\right\} \Rightarrow AIC \\ +Profusion\,of\ Complexity \end{array}\right\} \Rightarrow ICOMP$$

in several ways using the additivity properties of information theory. We further base our developments on similar considerations to Rissanen [9] in his *final estimation criterion (FEC)* in estimation and model identification problems, as well as *Akaike's* [7] *AIC*, and its analytical extensions in Bozdogan [10].

In *AIC*, the compromise takes place between the maximized log likelihood, i.e., $-2\log L(\hat{\theta}_k)$ (the *lack of fit component*) and *m*(*k*), the *number of free parameters* estimated within the model (the *penalty component*) which is a measure of complexity that compensates for the *bias* in the lack of fit when the *maximum likelihood estimators* (*MLEs*) are used. On the other hand, in *ICOMP*, we have a third term in the loss function

which is called the *"Profusion of Complexity"* which measures how the parameter estimates are correlated with one another in the model fitting process.

Therefore, instead of penalizing the number of free parameters directly, *ICOMP* penalizes the covariance complexity of the model. It is defined by

$$ICOMP = -2\log L(\hat{\theta}_k) + 2C(\hat{\Sigma}_{Model}), \tag{2}$$

where $L(\hat{\theta}_k)$ is the maximized likelihood function, $\hat{\theta}_k$ is the maximum likelihood estimate of the parameter vector θ_k under the model M_k, and C represents a real-valued complexity measure and $\hat{\Sigma}_{Model} = \hat{C}ov(\hat{\theta}_k)$ represents the estimated covariance matrix of the parameter vector of the model. This covariance matrix in *ICOMP* is estimated several ways. One of the ways to estimate this covariance matrix is to use the celebrated *Cramer-Rao lower bound (CRLB)* matrix through its inverse. That is, the *estimated inverse Fisher information matrix* (*IFIM*) $\hat{\mathcal{F}}^{-1}$ of the model given by

$$\hat{\mathcal{F}}^{-1} = \left\{ -E\left(\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'} \right)_{\hat{\theta}} \right\}^{-1}, \tag{3}$$

where the expression in bracket is the $(s \times s)$ matrix of second partial derivatives of the log-likelihood function of the fitted model evaluated at the maximum likelihood estimators $\hat{\theta}$. For this, see, Cramer [11] and Rao [12-14].

The estimated *IFIM* provides us an achievable accuracy of the parameter estimates by considering the entire parameter space of the model. *IFIM* is a measure of the best precision with which a parameter can be estimated from statistical data. It measures the quantum of information and measures the curvature of the log likelihood function of the model. The diagonal elements of *IFIM* contain the *estimated variances* or *squared standard errors* of the estimated parameters, while the off-diagonals of the matrix contain their covariances. When Akaike [7] was defining the accuracy of the parameter estimates by a universal criterion, he had *IFIM* in mind, but in his original derivation of *AIC*, he was not successful in bringing *IFIM* in the penalty term of his *AIC*, mathematically.

In its general form, for univariate and multivariate models (linear *and/or* nonlinear) *ICOMP* is defined by

$$ICOMP = -2\log L(\hat{\theta}_k) + 2C_1(\hat{\mathcal{F}}^{-1}), \tag{4}$$

where

$$C_1(\hat{\mathcal{F}}^{-1}) = \frac{s}{2}\log\left[\frac{tr\hat{\mathcal{F}}^{-1}}{s} \right] - \frac{1}{2}\log\left|\hat{\mathcal{F}}^{-1}\right| \tag{5}$$

is the maximal information complexity of the estimated inverse Fisher information matrix (IFIM) of the model, and where $s = \dim(\hat{\mathcal{F}}^{-1}) = rank(\hat{\mathcal{F}}^{-1})$. More on this later.

Hence, *ICOMP* in its idealized form is an additive composition of a term which measures the *lack of fit* (i.e., *inference uncertainty*), a second term which measures the *complexity of the covariance matrix of the parameter estimates* of a model, which represents the *parametric uncertainty* of a model. It provides a more judicious penalty term and balances the *overfitting* and *underfitting risks* of a model than that of *AIC*. Indeed, this new approach provides an entropic general *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.

Before we show and discuss several forms of *ICOMP class of criteria*, we first introduce some background material to understand the information-theoretic concept of complexity of a covariance matrix. Therefore, the rest of the paper is organized as follows. In Section 2, we define the information measure of dependence in higher dimensions and present the results on the initial definition of the information-theoretic measure of covariance complexity and also the maximal covariance complexity. Later, we provide other forms of complexity measures which are geometric and scale-invariant. In Section 3, we introduce the several forms of *ICOMP class of criteria* for model selection. For space considerations, we restrict the detailed proofs and derivations of these criteria where it is appropriate. For more details, we will refer the readers to Bozdogan [1, 5, 6, 10], Bozdogan and Ueno [16], Bozdogan and Bearse [17], and Bozdogan and Haughton [15].

We illustrate the practical utility and the importance of this new class of model selection criteria by providing a real example on customer profiling and segmentation of the mobile phone customers using a novel *multi-class support vector machine-recursive feature elimination* (*MSVM-RFE*) method. This is presented in Section 4. Section 5 concludes the paper with some discussion.

2. Information Theoretic Measure of Dependence and Complexity

When we are given a high dimensional data matrix *X* of size $(n \times p)$, in *multivariate statistical modeling* and *data mining*, often *p-variables* interact in some fashion or another, and some variables (or sets of variables) influence and exert effects on other variables. These effects are reflected in terms of interactions of these variables and their dependencies upon one another. We call this the *dependence* and *non-independence* of the variables involved. Therefore, for a random vector, we define the complexity as follows.

Definition 2.1: *The complexity of a random vector is a measure of the interaction*

or the dependency between its components.

The more interaction or the dependency there is, the larger the complexity will be. So, a high degree of complexity implies high rates of computational effort and statistical data processing.

We shall use information theory to analyze the *dependence* or *non-independence* and measure the complexity of set of variables.

2.1. Information Measure of Dependence in High-Dimensions

We consider a continuous *p*-dimensional distribution with joint density function $f(\mathbf{x}) = f(x_1,...,x_p)$ and marginal density functions $f_j(x_j), j = 1,...,p.$ Following Kullback [18], and Harris [19], Theil and Feibig [20], and others, we define the *information measure of dependence* between the random variables $X_1,\ldots,X_p$ as follows:

$$
\begin{aligned}
I(\mathbf{x}) = I(x_1,...,x_p) &= E_f[\log \frac{f(x_1,...,x_p)}{f_1(x_1)\cdots f_p(x_p)}] \\
&= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f(x_1,...,x_p)\log \frac{f(x_1,...,x_p)}{f_1(x_1)\cdots f_p(x_p)} dx_1\cdots dx_p,
\end{aligned}
\tag{6}
$$

where I is the *Kullback-Leibler (KL)* [21] *information divergence against independence*. It is a *measure of expected dependency* among the component variables. In the literature, this is also known as the *mutual information* or the *information proper*.

What is compared here is the *joint distribution* $f(x_1,...,x_p)$ of the random variables to the product of their *marginal* distributions $f_1(x_1)f_2(x_2)\cdots f_p(x_p)$ under the assumption that they are independently distributed to the extent that the joint distribution differs from the distribution of the variables under the assumption that they are independent. Hence, this is a *measure of interdependence* among the variables.

The properties of the *KL information divergence* are as follows:

- $I(\mathbf{x}) \equiv I(x_1,...,x_p) \geq 0$ i.e., the expected mutual information is nonnegative.
- $I(\mathbf{x}) \equiv I(x_1,...,x_p) = 0$ if and only if $f(x_1,...,x_p) = f_1(x_1)\cdots f_p(x_p)$ for every *p*-tuple $(x_1,...,x_p)$, i.e., if and only if the random variables $x_1,...,x_p$ are mutually statistically independent. In this case the quotient in (6) is equal to unity, and its logarithm is then zero. If it is not zero, this implies a *dependency*.

We relate the *KL divergence* in (6) to Shannon's [22] *entropy* by the important identity

$$I(\mathbf{x}) \equiv I(x_1,...,x_p) = \sum_{j=1}^{p} H(x_j) - H(x_1,...,x_p), \tag{7}$$

where

$$H(x_j) = -E\left[\log f(x_j)\right] = -\int_{-\infty}^{+\infty} f(x_j)\log f(x_j)dx_j, \tag{8}$$

is the marginal entropy, and

$$\begin{aligned} H(x_1,...,x_p) &= -E\left[\log f(x)\right] \\ &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f(x_1,...,x_p)\log f(x_1,...,x_p)dx_1\cdots dx_p \end{aligned} \tag{9}$$

is the *global* or *joint entropy.*

Watanabe [23] calls (7) the *strength of structure* and a *measure of interdependence*. We note that (7) is the sum of the interactions in a system with $x_1,...,x_p$ as components, which we define to be the entropy complexity of that system. This is also called the *Shannon complexity* (see, Rissanen [24]). If there is 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, [23]).

The relation in (7) can easily be generalized to finding the interaction between any subset of variables.

2.2. Information-Theoretic Measure of Covariance Complexity

To define the information-theoretic measure of complexity of a multivariate distribution, let $f(\mathbf{x}) = f(x_1,...,x_p)$ be a *multivariate normal* (*Gaussian*) *density function* given by

$$\begin{aligned} f(\mathbf{x}) &= f(x_1,...,x_p) \\ &= (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\}, \end{aligned} \tag{10}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2,..., \mu_p)', -\infty < \mu_j < \infty, j = 1,2,..., p$ and $\Sigma > 0$ (positive definite)

As a short hand, we write

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma). \tag{11}$$

Then the joint entropy $H(\mathbf{x}) = H(x_1,...,x_p)$ from equation (9) for the case in which $\boldsymbol{\mu} = \mathbf{0}$ is given by

$$\begin{aligned} H(\mathbf{x}) = H(x_1,...,x_p) &= -\int_{\mathbf{R}^p} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbf{R}^p} f(\mathbf{x}) \left[\frac{p}{2} \log(2\pi) |\Sigma| + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right] d\mathbf{x} \\ &= \frac{p}{2} \log(2\pi) |\Sigma| + \frac{1}{2} tr\left[\int_{\mathbf{R}^p} f(\mathbf{x}) \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})' d\mathbf{x} \right]. \end{aligned} \tag{12}$$

Then, since $E[\mathbf{x}\mathbf{x}'] = \Sigma$, we have

$$\begin{aligned} H(\mathbf{x}) = H(x_1,...,x_p) &= \frac{p}{2} \log(2\pi) + \frac{p}{2} + \frac{1}{2} \log |\Sigma| \\ &= \frac{p}{2} [\log(2\pi) + 1] + \frac{1}{2} \log |\Sigma|. \end{aligned} \tag{13}$$

See, e.g., Blahut [25].

Similarly, the marginal entropy $H(x_j)$ is

$$\begin{aligned} H(x_j) &= -\int_{-\infty}^{+\infty} f(x_j) \log f(x_j) dx_j \\ &= \frac{1}{2} \log(2\pi) + \frac{1}{2} + \frac{1}{2} \log(\sigma_j^2), j = 1,2,..., p \end{aligned} \tag{14}$$

where σ_j^2 is the variance of the j^{th} variable.

Using Shannon's [22] result, we have the following theorem.

Theorem 2.1: Among all distributions with the same mean and covariance matrix Σ, a *multivariate normal* (or *Gaussian*) distribution with that covariance matrix Σ has maximal entropy.

The proof of this theorem is given in Rao [26].

We note that the expression in (13) is a *fixed upper bound* for any *p-dimensional entropy* given the *covariance matrix*.

Since in general there is no unique distribution for which the maximum of entropy is achieved, a less restrictive result is given in van Emden [8]. This leads to a uniquely determined maximizing distribution due to the fact that not all covariances σ_{ij} are necessary to specify the entropy. This is already done by the determinant of Σ, and therefore, Shannon's [22] condition can be relaxed.

2.3. Initial Definition of Covariance Complexity

Van Emden [8] provides a reasonable initial definition of complexity of a covariance matrix Σ for the multivariate normal (or Gaussian) distribution. Using (7), this measure is given by:

$$\begin{aligned} I(x_1,...,x_p) \equiv C_0(\Sigma) &= \sum_{j=1}^{p} H(x_j) - H(x_1,...,x_p) \\ &= \sum_{j=1}^{p}\left[\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\sigma_{jj}) + \frac{1}{2}\right] - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{p}{2}. \end{aligned} \tag{15}$$

This reduces to

$$\begin{aligned} C_0(\Sigma) &= \frac{1}{2}\sum_{j=1}^{p}\log(\sigma_{jj}) - \frac{1}{2}\log|\Sigma| \\ &= \frac{1}{2}\log|Diag(\Sigma)| - \frac{1}{2}\log|\Sigma|, \end{aligned} \tag{16}$$

where $\sigma_{jj} \equiv \sigma_j^2$, is the variance of the j^{th} variable, and is the j^{th} diagonal element of Σ. The first term of (16) is not invariant under orthonormal transformations. As pointed out by van Emden [8], the result in (16) is not an effective measure of the amount of complexity in the covariance matrix Σ, since:

(i) $C_0(\Sigma)$ depends on the coordinates of the original random variables $x_1,...,x_p$.

(ii) The first term of $C_0(\Sigma)$ in (16) would change under orthonormal transformations.

2.4. 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 $x_1,...,x_p$ associated with the variances $\sigma_j^2,\ j = 1,2,...,p.$ As it is $C_0(\Sigma)$ in (16) 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 (16).

We do this by recognizing the fact that the maximum amount of complexity in Σ given by $C_0(\Sigma)$ under orthogonal transformations $y = Tx$ of variables, where $T'T = I$ and T is $(p \times p)$, may reasonably serve as a measure of complexity of Σ. This corresponds to observing the interaction between the variables under the coordinate system that most clearly represents it in terms of a measure $I(x_1,...,x_p) \equiv C_0(\Sigma)$. So, to improve upon $C_0(\Sigma)$ in (16), we seek the maximum of complexity in Σ under orthogonal transformations and have the following proposition.

Definition 2.1: A maximal information theoretic measure of complexity of a covariance matrix Σ of a multivariate Gaussian distribution is

$$\begin{aligned} C_1(\Sigma) &= \max_T C_0(\Sigma) = \max_T \{H(x_1) + \cdots + H(x_p) - H(x_1,...,x_p)\} \\ &= \frac{p}{2}\log\left[\frac{tr(\Sigma)}{p}\right] - \frac{1}{2}\log|\Sigma|, \end{aligned} \tag{17}$$

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

Hence, based on the $C_0(\Sigma)$ measure, the idea of using the maximal information complexity measure $C_1(\Sigma)$ as a penalty functional, in the literature, is due to Bozdogan [1, 2, 5, 6], where the proof and the properties of $C_1(\Sigma)$ are shown in detail. For this, see, also, Mulaik [28] in his interesting book on *Linear Causal Modeling with Structural Equations*. For space considerations, we do not recapitulate these proofs here.

If we let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the eigenvalues of the covariance matrix Σ then

$$\bar{\lambda}_a = tr(\Sigma)/\ p = 1/\ p\sum_{j=1}^{p}\lambda_j$$

is the *arithmetic mean of the eigenvalues*, and

$$\bar{\lambda}_g = |\Sigma|^{1/p} = \left(\prod_{j=1}^{p}\lambda_j\right)^{1/p}$$

is the *geometric mean of the eigenvalues* of Σ. Then the complexity of Σ can be written as

$$C_1(\Sigma) = \frac{p}{2}\log\left(\bar{\lambda}_a / \bar{\lambda}_g\right). \tag{18}$$

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 [28] in his paper shows the fact that the *trace* (*sum of variances*) and the *determinant* of the

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

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.

Other interpretations of $C_1(\Sigma)$ are given in Bozdogan [1, 2, 5, 6].

Although $C_1(\Sigma)$ in (17) is derived under the existence of an orthogonal transformation, we prefer to use this over the initial definition of complexity $C_0(\Sigma)$ in (16) without using any transformations of Σ. This is because of its rather attractive properties. We do not necessarily discard the use of $C_0(\Sigma)$. Further, $C_1(\Sigma)$ is monotonically increasing function of the dimension *p* of Σ. Compared to $C_0(\Sigma)$ defined in (16), $C_1(\Sigma)$ is a preferred measure to evaluate complexity which is much less costly to compute in higher dimensions.

2.5. Percent Relative Reduction in Complexity

Diagonal operation of a covariance matrix Σ always reduces the complexity of Σ. Let P be the correlation matrix obtained from Σ by the relationship $P = D_\sigma^{-1/2}\Sigma D_\sigma^{-1/2}$, where $D_\sigma = Diag(1/\sigma_1,\ldots,1/\sigma_p)$ is a diagonal matrix whose diagonal elements equals $1/\sigma_j,\ j = 1,\ldots,p$. From (17), we have

$$C_1(P) = -\frac{1}{2}\log|P| \equiv C_0(P) \tag{19}$$

and $C_1(P) \equiv C_0(P)$ takes into account the *interdependencies (correlations)* among the variables. Then, the *relative reduction of complexity* is given by

$$RRC = \frac{C_1(\Sigma) - C_1(P)}{C_1(\Sigma)}. \tag{20}$$

Percent relative reduction of complexity is then

$$PRRC = \frac{C_1(\Sigma) - C_1(P)}{C_1(\Sigma)} \times 100\%. \tag{21}$$

For simplicity, the $C_0(P)$ measure based on *correlation matrix* will be denoted by C_R and $C_0(P)$ is written as C_R for notational convenience. 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 $|P| = 1$, then $I(x_1,\ldots,x_p) \equiv C_0(\Sigma) = 0$ which implies the mutual independence of the variables $x_1,\ldots,x_p$. If the variables are not mutually independent, then $0 < |P| < 1$ and that

$I(x_1,...,x_p) \equiv C_0(\Sigma) > 0$.In this sense, the information measure of dependence, $I(x_1,...,x_p)$ can also be viewed as a measure of *dimensionality of model manifolds*.

2.6. Other Forms of Complexity

Under the orthogonal transformation *T*, the maximal complexity in (17) can be written as

$$\begin{aligned} C_1^*(\Sigma) &= -\frac{1}{2}\sum_{j=1}^{p}\log(s\lambda_j) \\ &\cong \frac{1}{4}\sum_{j=1}^{q}(s\lambda_j - 1)(s\lambda_j - 3) - \frac{1}{6}\sum_{j=1}^{q}O(s\lambda_j - 1)^3. \\ 0 < \lambda_j &< \frac{2}{s},\ j = 1,2,\ldots,s \end{aligned} \tag{22}$$

where "$\log$" denotes the natural logarithm "$\ln$", $O(\bullet)$ denotes the order of the argument. The Taylor expansion of $\log(s\lambda_j)$ used in (22) is about the neighborhood of the point

$$\lambda_1 = \lambda_2 = \cdots = \lambda_s = \frac{1}{s}. \tag{23}$$

At the point of eigenvalue equality $C_1^*(\Sigma) = 0$, with $C_1^*(\bullet) > 0$ otherwise. See, Morgera [29].

As is explained in van Emden [8], we note that (22) is only one possible measure of covariance complexity. Any convex function $\phi(\bullet)$, like $-\ln(\bullet)$, whose second derivative exists and is positive, may be used as a complexity measure, i.e.,

$$C_\phi^*(\bullet) = c\sum_{j=1}^{q}\left[\phi(\lambda_j) - \phi(\frac{1}{q})\right] \tag{24}$$

leads to an entire family of complexity measures, where c is a constant. For more on convex functions, see Conway [30].

2.6.1. Frobenius Norm Complexity

With the convexity in mind, van Emden [8] suggested a second measure of complexity of a covariance matrix based on the Frobenius norm given by

$$C_F(\Sigma) = \frac{1}{p}\|\Sigma\|^2 - \left(\frac{tr\Sigma}{p}\right)^2, \tag{25}$$

where $\|\Sigma\|^2 = tr(\Sigma'\Sigma)$, the square of the Frobenius norm of Σ which is invariant under orthogonal transformations.

In terms of the *eigenvalues* (or *singular values*), $C_F(\Sigma)$ reduces to

$$C_F(\Sigma) = \frac{1}{p}\sum_{j=1}^{p}(\lambda_j - \bar{\lambda}_a)^2, \tag{26}$$

where p is the rank of Σ, λ_j is the j^{th} eigenvalue of $\Sigma > 0$, $j = 1,2,\ldots,p$ and $\bar{\lambda}_a$ is arithmetic mean of the eigenvalues. Note that $C_F(\Sigma) \geq 0$ with $C_F(\Sigma) = 0$ only when all $\lambda_j = \bar{\lambda}_a$ .Hence, $C_F(\Sigma)$ measures the *absolute variation in the eigenvalues* and it is translation invariant. That is, $C_F(\Sigma + kI) = C_F(\Sigma)$. But it is not scale-invariant.

Also, note that $C_F(\Sigma)$ is convenient to compute since no transformation of Σ is required and it is applicable to any covariance matrix (van Emden [8] and Morgera [29]).
If we define the spread of the covariance matrix Σ by

$$s(\Sigma) = \max_{j,k} \left| \lambda_j - \lambda_k \right|, \tag{27}$$

then we can obtain an *upper* and *lower bound* on $C_F(\Sigma)$ given by

$$\frac{1}{2p} s^2(\Sigma) \leq C_F(\Sigma) \leq \frac{p(p-1)}{2p^2} s^2(\Sigma), \tag{28}$$

where $s^2(\Sigma)$ denotes the square of the spread in (27), and p is the rank or the dimension of Σ, i.e., the number of variables.

2.6.2. Scale-Invariant Complexity

The connection between the maximal information complexity $C_1(\Sigma)$ and Frobenius norm complexity $C_F(\Sigma)$ is that these two complexities are *second order equivalent*. The proof of this is given in van Emden [8] by an incorrect sign. But the corrected sign is given in Morgera [29] in the power series expansion of $\log(x)$. Hence, we can approximate $C_1(\Sigma)$ in terms of the eigenvalues λ_j, $j = 1,2,\ldots,p$ by

$$C_1(\Sigma) \cong \frac{1}{4} \sum_{j=1}^{p} \left(\frac{\lambda_j - \bar{\lambda}_a}{\bar{\lambda}_a} \right)^2. \tag{29}$$

Now, we can relate $C_1(\Sigma)$ to the Frobenius norm characterization of complexity $C_F(\Sigma)$ of Σ (Bozdogan, [1]) by introducing $C_{1F}(\Sigma)$ given by

$$\begin{aligned} C_{1F}(\Sigma) &= \frac{p}{4} \frac{C_F(\Sigma)}{\left(\frac{tr(\Sigma)}{p} \right)^2} = \frac{p}{4} \frac{\frac{1}{p} \| \Sigma \|^2 - \left(\frac{tr(\Sigma)}{p} \right)^2}{\left(\frac{tr(\Sigma)}{p} \right)^2} \\ &= \frac{p}{4} \frac{\frac{1}{p} tr(\Sigma'\Sigma) - \left(\frac{tr(\Sigma)}{p} \right)^2}{\left(\frac{tr(\Sigma)}{p} \right)^2}. \end{aligned} \tag{30}$$

In terms of the eigenvalues, $C_{1F}(\Sigma)$ becomes

$$C_{1F}(\Sigma) = \frac{p}{4}\frac{1}{\mathrm{p}\bar{\lambda}_a^2}\sum_{j=1}^{p}(\lambda_j - \bar{\lambda}_a)^2 = \frac{1}{4\bar{\lambda}_a^2}\sum_{j=1}^{p}(\lambda_j - \bar{\lambda}_a)^2. \tag{31}$$

We note that $C_{1F}(\Sigma)$ is a *second order equivalent measure of complexity* to the original $C_1(\Sigma)$ measure. Also, we note that $C_{1F}(\Sigma)$ is *scale-invariant* and $C_{1F}(\Sigma) \geq 0$ with $C_{1F}(\Sigma) = 0$ only when all when all $\lambda_j = \bar{\lambda}_a$. Also, $C_{1F}(\Sigma)$ measures the *relative variation in the eigenvalues* rather than *absolute variation of the eigenvalues*.

In summary, we observe that $C_1(\Sigma)$ and $C_{1F}(\Sigma)$ are quite different measures of complexity of Σ. In the literature, several authors including Rissanen [9], Ljung & Rissanen [31], Maklad & Nichols [32], and Morgera [29], have made use of van Emden's [8] results, all with in some form of incomplete arguments without the *lack of fit part* as a model selection index. Further, Poskitt [33] also used van Emden's [8] results in trying to discriminate Bayesian models with an error. These incomplete arguments in the earlier contributions that are cited here are rectified in this authors work and they are generalized through this author's several unique scientific contributions.

Next, we introduce several forms of information-theoretic measure of complexity criterion called *ICOMP (IFIM)* of Bozdogan [1-6] as a decision rule for model selection evaluation based on the *maximal covariance complexity* $C_1(\bullet)$, $C_R(\bullet)$, and $C_{1F}(\bullet)$. These approaches have established theoretical background and foundation. Again, for space limitations, here we only show selectively the derivations of certain forms and refer the readers to the other forms which are already established in the literature.

3. A New Class of Information Complexity (ICOMP) Criteria

In this section, we introduce several forms of *ICOMP class of criteria* for model selection 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* discussed in detail in Section 2.

3.1. ICOMP as an Approximation to the Sum of Two Kullback-Leibler Distances

Here, we first introduce a general formulation of *ICOMP* using the estimated *inverse-Fisher information matrix* (*IFIM*), which is also known as the *Cramer-Rao lower bound* (*CRLB*) *matrix*. This approach to *ICOMP* is an approximation to the *sum of two Kullback-Leibler* (*KL*) [21] distances. Such an approach provides us an achievable accuracy of the parameter estimates by considering the entire parameter space of the model. As a result we have:

Proposition 3.1: *For a multivariate normal linear or nonlinear model we define thegeneral form of ICOMP (IFIM) as*

$$ICOMP\left(IFIM\right) = -2\log L\left(\hat{\theta}_M\right) + 2C_1\left(\hat{\mathcal{F}}^{-1}\right), \tag{32}$$

where C_1 denotes the *maximal informational complexity* of $\hat{\mathcal{F}}^{-1}$, the estimated *IFIM* given by

$$C_1(\hat{\mathcal{F}}^{-1}) = \frac{s}{2}\log\left[\frac{tr\hat{\mathcal{F}}^{-1}}{s}\right] - \frac{1}{2}\log\left|\hat{\mathcal{F}}^{-1}\right|, \tag{33}$$

where $s = \dim(\hat{\mathcal{F}}^{-1}) = rank(\hat{\mathcal{F}}^{-1})$.

For detailed derivations, see Bozdogan and Haughton [18], Bearse and Bozdogan [17], Bozdogan [4, 5].

The first component of *ICOMP (IFIM)* in (32) measures the *lack of fit of the model*, and the second component measures the complexity of the *estimated inverse- Fisher information matrix* (*IFIM*), which gives a scalar measure of the celebrated *Cramer-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. It is an intrinsic measure of uncertainty, and, furthermore, it is a quality metric of the estimation procedure. For more on this and for some immediate physical motivation, we refer the readers to the interesting book by Frieden [34], entitled: *"Physics from Fisher Information."*

The use of $C_1\left(\hat{\mathcal{F}}^{-1}\right)$ 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.

We note that, 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* given by

$$ICOMP\left(IFIM\right) = -2\log L\left(\hat{\theta}_M\right) + s\log\left(\bar{\lambda}_a / \bar{\lambda}_g\right), \tag{34}$$

where $s = \dim \hat{\mathcal{F}}^{-1}\left(\hat{\theta}\right) = rank\hat{\mathcal{F}}^{-1}\left(\hat{\theta}\right)$, and where $\bar{\lambda}_a$ is the arithmetic mean and $\bar{\lambda}_g$ is the geometric mean of the eigenvalues of $\hat{\mathcal{F}}^{-1}$.

We note that *ICOMP (IFIM)* now looks in appearance like the *CAIC* of Bozdogan [10], Rissanen's [35] *MDL*, and Schwarz's [36] Bayesian criterion *SBC*, except for using $\log\left(\bar{\lambda}_a / \bar{\lambda}_g\right)$ instead of using $\log(n)$ denotes the natural logarithm of the sample size n.

A model with minimum *ICOMP (IFIM)* is chosen to be the best among all possible competing alternative models.

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, MDL, SBC, or CAIC*. The lack of parsimony and the profusion of complexity are automatically adjusted by $C_1\left(\hat{\mathcal{F}}^{-1}\right)$ across the competing alternative portfolio of models as the parameter spaces of these models are constrained in the model selection process.

In the literature, several authors such as McQuarie and Tsai [37], Burnham and Anderson [38], and a few others, without reviewing underlying statistical theory and the impact of *ICOMP (IFIM)* in terms of its potential and innovation, have erroneously interpreted the contribution of this novel approach over *AIC*, and *AIC-type* criteria. We believe that this is due to of not being able to compute *IFIM* correctly under various multivariate models.

3.2. ICOMP as an Estimate of Posterior Expected Utility

The idea of using two utility functions U_1 and U_2 that are multiplied to define a utility U whose *posterior expectation* is (approximately) maximized to select a model was considered notably by Poskitt [33], and others. If we relate utility U_1 to the *lack of fit component* of the model and U_2 to the *complexity of the parameter space of the model,* i.e., *the dimension of the model*, we introduce a new *ICOMP* class of criteria as a Bayesian criterion in maximizing a *posterior expected utility (PEU)* following the results from Bozdogan and Haughton [15].

Proposition 3.2: *ICOMP as a Bayesian criterion in maximizing a posterior expected utility (PEU) is given by*

$$ICOMP\left(IFIM\right)_{PEU} = -2\log L\left(\hat{\theta}_M\right) + k + 2C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right). \tag{35}$$

Proof:

Let $L_M\left(y,\theta\right)$ be the likelihood function of the parameter vector for a given vector y of observations. Let $f_{Prior}\left(\theta \mid M\right)$ denote the prior density function of θ on the model M, and let $f_{Post}\left(\theta \mid y\right)$ be the posterior density of θ corresponding to the prior $f_{Prior}\left(\theta \mid M\right)$. Let $FIM\left(\theta\right)$ denote the *Fisher information matrix* (*FIM*) for the n observations corresponding to model M, and let k be the dimension of M. Following Poskitt [33], we consider the KL distance between the *posterior* and the *prior densities* for model M given by

$$\begin{aligned} KL\left(f_{Post}\left(\theta \mid y\right); f_{Prior}\left(\theta \mid M\right)\right) &= \int_{\Theta_M} f_{Post}\left(\theta \mid y\right)\log f_{Post}\left(\theta \mid y\right)d\theta \\ &\quad - \int_{\Theta_M} f_{Post}\left(\theta \mid y\right)\log f_{Prior}\left(\theta \mid M\right)d\theta \\ &= H\left(f_{Post}\left(\theta \mid y\right)\right) - \int_{\Theta_M} f_{Post}\left(\theta \mid y\right)\log f_{Prior}\left(\theta \mid M\right)d\theta. \end{aligned} \tag{36}$$

Further following the arguments in Poskitt [33], under regularity conditions which guarantee the asymptotic normality of the posterior distribution, that is, when

$$f_{Post}\left(\theta \mid y\right) \cong N\left(\hat{\theta}, \Sigma\left(\hat{\theta}\right) \equiv \hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right) \tag{37}$$

$$KL\left(f_{Post}\left(\theta \mid y\right); f_{Prior}\left(\theta \mid M\right)\right) \cong -\frac{k}{2}\log\left(2\pi\right) - \frac{k}{2} - \frac{1}{2}\log\left|\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right| - \log f_{Prior}\left(\theta \mid M\right). \tag{38}$$

One can argue, as Poskitt [33] that a utility U_1 can be defined as $\log(U_1) = KL$ given by (38). In Bayesian design of experiments, following Lindley's [39] suggestion, several authors have considered *KL* as a utility function. For more on this, see, e.g., Chaloner and Verdinelli [40].

In our case, we propose to multiply the utility U_1 by a utility U_2 equal to:

$$U_2 = \exp\left[-C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right)\right] \tag{39}$$

Then our utility $U = U_1 \times U_2$, and the log of that utility is

$$\log U = \log U_1 + \log U_2. \tag{40}$$

Hence, substituting (38) and (39) into (40), we have

$$\log(U) = -\frac{k}{2}\log(2\pi) - \frac{k}{2} - \frac{1}{2}\log\left|\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right| - \log f_{Prior}(\theta \mid M) - C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right) \tag{41}$$

which is the difference of *KL* distances. Note that our utility U_2 is slightly different from that used by Poskitt. His utility U_2 uses only the *trace term* in the expression of the complexity, and does not contrast the determinant of *IFIM* with the trace. The trace involves only the diagonal elements analogous to *variances* while the determinant involves also the off diagonal elements analogous to *covariances*. This amounts to a comparison of the *geometric* and *arithmetic means of the eigenvalues* of *IFIM* given in (34).

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. If we apply Poskitt's *Corollary 2.2* or the Laplace expansion results of Kass, Tierney and Kadane [41], it follows that, under some regularity conditions, if the parameter vector θ lies in M, the log of the *posterior expected utility*(*PEU*) can be approximated by

$$\begin{aligned}\log(PEU) \cong \log f\left(y, \hat{\theta}_M\right) + \frac{k}{2}\log(2\pi) + \frac{1}{2}\log\left|\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right| + \log(U) \\ + \log f_{Prior}\left(\hat{\theta}_M \mid M\right) + \log f(M)\end{aligned} \tag{42}$$

up to order $O(1/n)$ and up to some terms which do not depend on the model M. Replacing $\log(U)$ in (42) by its value in (41), and simplifying, some terms will cancel out. We thus obtain a criterion, to be *maximized to choose a model*, equal to:

$$\log f\left(y, \hat{\theta}_M\right) - \frac{k}{2} - C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right) + \log f(M) \tag{43}$$

Maximizing (43) is equivalent to minimizing estimated *ICOMP (IFIM)* given by

$$ICOMP(IFIM)_{PEU} = -2\log L\left(\hat{\theta}_M\right) + k + 2C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right) + \log f(M) \tag{44}$$

Assuming that the prior on model M, that is, $f(M)$, is constant for all models in (44), we have

$$ICOMP\left(IFIM\right)_{PEU} = -2\log L\left(\hat{\theta}_M\right) + k + 2C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right). \tag{45}$$

This completes the derivation of $ICOMP\left(IFIM\right)_{PEU}$.

The decision rule is to choose the minimum of *ICOMP (IFIM)* over the class of models $M_k,\ k = 1,2,\ldots,K$ that is the best fitting model.

We note that *ICOMP (IFIM)* in (45) penalizes the bad scaling of the parameters. If scale invariance is an issue in model selection enterprise, one can use the *correlational form* of *IFIM* given by

$$\hat{\mathcal{F}}_R^{-1}\left(\hat{\theta}_M\right) = D_{\hat{\mathcal{F}}^{-1}}^{-1/2}\hat{\mathcal{F}}^{-1}D_{\hat{\mathcal{F}}^{-1}}^{-1/2}, \tag{46}$$

where $D_{\hat{\mathcal{F}}^{-1}}^{-1/2}$ is the negative square-root of the diagonal entries of $\hat{\mathcal{F}}^{-1}$. Hence, the correlational form of $ICOMP\left(IFIM\right)_{PEU}$ is:

$$ICOMP\left(IFIM\right)_{R\_PEU} = -2\log L\left(\hat{\theta}_M\right) + k + 2C_1\left(\hat{\mathcal{F}}_R^{-1}\left(\hat{\theta}_M\right)\right). \tag{47}$$

In nonlinear modeling, when the parameter estimates are highly correlated, one can remove the correlation by considering parameter transformations of the model. The difference between the complexities $C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right)$ and $C_1\left(\hat{\mathcal{F}}_R^{-1}\left(\hat{\theta}_M\right)\right)$ can be used to show how well the parameters are scaled. 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 the scaling effect. Indeed, the reduction in the correlation may even make scaling worse. In this sense, *ICOMP (IFIM)* may be better than $ICOMP\left(IFIM\right)_R$ in model selection, since it considers both of these effects in one criterion function. For more on this, see, e.g., Chen [42] in his doctoral thesis under the supervision of this author.

3.3. Other and Consistent Forms of $ICOMP_{PEU}$

Note that when we defined the utility

$$U_2 = \exp\left[-a \times C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right)\right], \tag{48}$$

we considered the constant multiplier $a \equiv 1$ in the exponent of U_2 to obtain the result in (45). This formulation, gives us the additional term *k* in the penalty function which is the number of estimated parameters in the model M .

Indeed other choices of a and the utility U_2 are possible and equally justifiable giving rise to different penalty functions. For example, a choice of $a \equiv \log\left(n\right)$, would yield

$$ICOMP\left(IFIM\right)_{PEU\_LN} = -2\log L\left(\hat{\theta}_M\right) + k + 2\log\left(n\right)C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right). \tag{49}$$

Different choices of utility U_2 may depend on other characteristics that a researcher can consider on the parameter vector θ_M if the model M is under consideration. Therefore, the full specification of the form of the utility function U_2 is important. By defining different forms of the utility U_2 we can, therefore, obtain other forms of $ICOMP\left(IFIM\right)_{PEU}$ that give us many useful class of model selection criteria.

These are given as follows.

- The choice of the utility

$$U_2 = \exp\left[-tr\left(\hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\right) - C_1\left(\hat{\mathcal{F}}^{-1}\right)\right] \tag{50}$$

would lead to

$$\begin{aligned} ICOMP\left(IFIM\right)_{PEU\_Miss} &= -2\log L\left(\hat{\theta}_M\right) + k + 2tr\left(\hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\right) + 2C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right) \\ &= -2\log L\left(\hat{\theta}_M\right) + k + 2\left[tr\left(\hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\right) + C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right)\right]. \end{aligned} \tag{51}$$

We can approximate $tr\left(\hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\right)$ by

$$tr\left(\hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\right) \cong \frac{nk}{n-k-2} \tag{52}$$

which corrects the *bias* for small as well as large sample sizes if the model is misspecified. Note that $tr\left(\hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\right)$ is the well-known Lagrange-multiplier test statistic. See, for example, Takeuchi [43], Hosking [44], and Shibata [45].

If the model is correctly specified, then

$$tr\left(\hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\right) = tr(I_k) = k. \tag{53}$$

Therefore, $ICOMP\left(IFIM\right)_{PEU\_Miss}$ reduces to

$$\begin{aligned} ICOMP\left(IFIM\right)_{PEU\_AIC_3} &= -2\log L\left(\hat{\theta}_M\right) + k + 2k + 2C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right) \\ &= -2\log L\left(\hat{\theta}_M\right) + 3k + 2C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right) \\ &= AIC_3 + 2C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right). \end{aligned} \tag{54}$$

Note that in utility U_2 above in (50), the terms in the exponent can be also be weighted differently. There is no necessity that these terms be weighted equally.

- The choice of the utility

$$U_2 = \exp\left[-\frac{k}{2}\log(n) - C_1\left(\hat{\mathcal{F}}^{-1}\right)\right] \tag{55}$$

would lead to

$$\begin{aligned} CICOMP\left(IFIM\right)_{PEU\_CAIC} &= -2\log L\left(\hat{\theta}_M\right) + k + k\log(n) + 2C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right) \\ &= CAIC + 2C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right). \end{aligned} \tag{56}$$

As we can see, by choosing different forms of utility U_2, we can obtain several other interesting forms of *ICOMP class of criteria* that are consistent and penalize overparameterization more stringently to pick only the simplest models whenever there is nothing to be lost by doing so.

3.4. ICOMP for Misspecified Models

In this section we generalize *ICOMP* to the case of a misspecified model and develop *ICOMP* under misspecification. Suppose that the fitted model is the wrong or misspecified model. Then we have

Proposition 3.3: *Under model misspecification, ICOMP is defined by*

$$\begin{aligned} ICOMP\left(Model\right)_{Misspec} &= -2\log L\left(\hat{\theta}\right) + 2C_1\left(\widehat{Cov}\left(\hat{\theta}\right)_{Misspec}\right) \\ &= -2\log L\left(\hat{\theta}\right) + 2C_1\left(\hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\hat{\mathcal{F}}^{-1}\right). \end{aligned} \tag{57}$$

Or, equivalently

$$ICOMP\left(IFIM\right)_{PEU\_Miss} = -2\log L\left(\hat{\theta}_M\right) + k + 2\left[tr\left(\hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\right) + C_1\left(\hat{\mathcal{F}}^{-1}\left(\hat{\theta}_M\right)\right)\right]. \tag{58}$$

When a model is misspecified the *"sandwich"* or *"robust"* covariance matrix consistently estimated by

$$\hat{C}ov(\hat{\theta})_{Misspec} = \hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}\hat{\mathcal{F}}^{-1}. \tag{59}$$

If the model is correct, we get

$$\mathcal{F} = \mathcal{R} \tag{60}$$

and the *"sandwich"* or *"robust"* covariance matrix reduces to

$$Cov\left(\theta\right) = \mathcal{F}^{-1} \tag{61}$$

the usual *inverse Fisher information matrix* (*IFIM*), which is known as the *'naive'* covariance formula.

3.5. ICOMP as a Performance Measure: $ICOMP_{PERF}$

In many classification and clustering problems and when we use kernel-based methods such as the *support vector machines* (*SVM*), *Multi-Class SVM* (*MSVM*), etc., our goal is to minimize the probability of misclassification error. Intuitively, then, the penalty term for a poorly-fitting model would be based on the classification error rate. In *SVM* and *MSVM*

type of problems, the error variance σ^2 is estimated by the *mean squared difference between actual group labels* (y_i) and *predicted group labels* ($\hat{y}_i$) given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{62}$$

Now following the work of Howe and Bozdogan [46] the information-theoretic measure of complexity as performance measure is defined as follows

$$ICOMP_{PERF} = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2C_{1F}(\hat{\Sigma}_{STA\_CSE}) \tag{63}$$

where $\hat{\Sigma}_{STA\_CSE}$ is the *stabilized and smoothed convex sum covariance matrix estimator* (*STA-CSE*) given by

$$\hat{\Sigma}_{STA\_CSE} = \frac{n}{n+m}\hat{\Sigma}_{STA} + (1 - \frac{n}{n+m})\hat{D}_{STA}, \tag{64}$$

where

$$\hat{D}_{STA} = \left(\frac{1}{p}tr(\hat{\Sigma}_{STA})\right) I_p \tag{65}$$

and

$$C_{1F}(\hat{\Sigma}_{STA\_CSE}) = \frac{1}{4\bar{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \bar{\lambda}_a)^2. \tag{66}$$

First, the hybrid covariance estimate in (64) is calculated, and then the diagonal matrix of the largest singular values as a reduced rank approximation of $\hat{\Sigma}_{STA\_CSE}$ is obtained. By minimizing $ICOMP_{PERF}$ the classification error is minimized under the best fitting model. Also, $ICOMP_{PERF}$ is used to choose an optimal kernel function in kernel-based methods. We also use $ICOMP_{PERF}$ in *SVM-RFE* subset selection problems.

In the literature cross-validation-based criteria has been used for variable selection. These types of criteria are too time-consuming due to the high-dimensionality of the feature space. The proposed method shortens the variable selection time.

4. A Real Numerical Example: Customer Profiling and Segmentation

In this section, we show a real numerical example using a novel and flexible supervised classification technique called *Multiclass Support Vector Machines Recursive Feature Elimination* (*MSVM-RFE*) for segmentation of mobile phone customer data base and to identify the best features for customer profiling and management in *K* classes or groups. Description of a customer group or type of customer based on various *demographic*, *psychographic* and/or *geographic characteristics*, is called *customer* or *shopper profile*. The characteristics of a customer may include *income*, *occupation*, *level of education*, *age*, *gender*, *hobbies*, or *area of residence*. Customer profiles provide the knowledge needed for a company to select the best prospect of customers that maximize the profitability of the company by establishing a one-to-one relationship with the customer.

First we give a brief set up of *binary support vector machine* (*SVM*) and *Multi-Class SVM* (*MSVM*).

4.1. Binary Support Vector Machines (SVMs)

Support vector machines (SVMs) are modern classification algorithms which are not sensitive to the curse of dimensionality and well suited for the analysis of high dimensional data. Intuitively, an *SVM* searches for a hyper plane with maximal distance between itself and the closest observation from each of the classes (Vapnik [47]). Therefore, *SVM is a maximum margin classifier* and the *decision function of SVMs is represented as a linear function in feature space as*

$$f(x_i) = \left\langle w^*, k_s(x_i) \right\rangle + b^*, \tag{67}$$

where $k_s(x_i) = \left[K(x_i, s_1), K(x_i, s_2), \ldots, K(x_i, s_m)\right]$ is the vector of the i^{th} data point evaluated at the m support vectors, which form a subset of the data. This is the *support vector machine (SVM)*. Thus, optimization of the weights w^* and intercept b^* becomes the *quadratic programming* (*QP*) problem given by

$$\begin{aligned} \left(w^*, b^*\right) &= \min_{w,b}\left[\frac{1}{2}\| w \|^2 + C\sum_{i=1}^{n}\xi^d\right], \\ \text{subject to } &\begin{cases} \left\langle w, x_i \right\rangle + b \geq 1 - \xi_i, \ i \in I_1, \\ \left\langle w, x_i \right\rangle + b \geq -1 + \xi_i, \ i \in I_2, \\ C > 0, \xi_i \geq 0, i \in I_1 \bigcup I_2 \end{cases} \end{aligned} \tag{68}$$

When $d = 1$, we say the *SVM* is L_1 *soft margin trained*, otherwise, it is L_2 *soft margin trained*. *C* is a regularization constant, and I_1 and I_2 are slack variables used to relax the inequalities for non-separable data.

4.2. Multi-Class Support Vector Machines (MSVMs)

For data composed of $K > 2$ classes or groups indexed by $k, k = 1, \ldots, K$, we consider a set of discriminant functions

$$f_k(x_i) = \left\langle w_k, k_s(x_i) \right\rangle + b_k. \tag{69}$$

There are several ways to decompose the MSVM, including *One-Against All (OAA)* and *One-Against-One (OAO)* - see Hsu and Lin [48].The *OAA* decomposition works by trading the single multi-class problem for *K* binary *SVM* problems, where the binary state vector y_k' is

$$\mathrm{y}_k' = \begin{cases} 1 & for \quad y = y_k \\ 2 & for \quad y \neq y_k \end{cases} \tag{70}$$

For example, if we had $K = 3$ classes A, B, and C, *OAA* would solve 3 binary problems:

$$\begin{aligned} A \quad vs \quad BC \\ B \quad vs \quad AC \\ C \quad vs \quad AB \end{aligned}$$

The multi-class classification rule used is then

$$q(x_i) = \max_{k=1,2,\ldots,K} f_k(x_i). \tag{71}$$

OAO, on the other hand, solves the multi-class problem by solving

$$K' = \frac{K(K-1)}{2}$$

binary SVM problems, in which all pairs of classes are considered. The majority voting strategy shown in equation (71) is used to select the final class assignments $vote(x_i) = [v_1(x_i), v_2(x_i), \ldots, v_{K'}(x_i)]$ is a vector indicating the frequency with which, from all K' *binary SVM* results, the i^{th} data point was classified into each group.

$$q(x_i) = \max_{y_i'=1,2,\ldots,K'} vote(x_i). \tag{72}$$

Using the same groups A, B, and C, *OAO* solves the *binary SVMs*

$$A \quad vs \quad B$$
$$A \quad vs \quad C$$
$$B \quad vs \quad C$$

We use and score $ICOMP_{PERF}$ defined in (63) which was inspired by the regression basis of discriminant analysis given by

$$ICOMP_{PERF} = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2C_{1F}(\hat{\Sigma}_{STA\_CSE}) \tag{73}$$

where

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{74}$$

is the estimated error variance between actual response values and predicted response values. That is, the *error variance term for a supervised classification*. $\hat{\Sigma}_{STA\_CSE}$ is the estimated hybridized stabilized and smoothed covariance matrix of the MSVMs.

4.3. The Description of the Dataset

We present our numerical results on customer profiling and segmentation of mobile phone customer data base. This dataset is the courtesy of Camillo [49]. Our goal is to:

- *Optimally classify the mobile phone customers in Italian market;*
- *Choose the best subset of Principal Component dimensions of the original customer profile variables; and*
- *Determine the best strategy for the Italian cell phone company.*

The customer data base is based on Italian mobile phone users of TELECOM. Presently, the company has about 95% market penetration rate in Italy in terms of its services. Customer data base was created from a survey study of $n = 1,021$ customers on 90 items that measured the lifestyle behavior of the customers, their knowledge of mobile phone technology and its usage, etc. After pre-processing and cleaning the data base and using optimal scaling on $n = 1,021$ customers, the 90 items were characterized on $d = 20$ principal components (PCs). So the database which is analyzed here consists of $n = 1,021$ customers (or observations) on $d = 20$ PC features with no missing observations and

$K = 4$ distinct groups. Details and the purpose on $K = 4$ groups are given in the report of Camillo, Liberati, and Athappilly [50]. These are:

- **Group 1:*"Functional People"***(20.8%) –Generally males 35-44 years old, with low education. They call and receive calls from people belonging to the same cell brand company. They use short message service (SMS) regularly. They use mobile phone especially for professional needs and choose the mobile phone based on its functional features.
- **Group 2:*"Practical People"***(35.9%)– Generally males and females of 45-64 years old, with medium/high education. They use phones especially in the afternoons (between 3:00-6:00 pm). They spend about 30 Euros per month and use all the services because they have an intense social life.
- **Group 3:"Techno People"**(27.59%)– Generally males of 25-34 years old and highly educated. They live in the North-West of Italy. They are for the most part professional men or students, who spend about 45 Euros per month. They buy high technology tools for their mobile phone. They also have an intense social life.
- **Group 4: *"Mature People"***(15.8%)–Generally over 65 years old females, who live mostly in the big cities in the North-West of Italy. They use the mobile phone only for emergencies. On other occasions, they prefer to call with home phone.

The goal of customer profiling and segmentation is to study the *customer relationship management* (*CRM*), which is to identify the best business customers-in terms of customer acquisition and retention; to choose the appropriate medium to reach the best customers prioritize the best business target markets for possible expansion; and to identify the best customer profiles for predictive data mining of future customers, etc.

4.4 Classical Discriminant Analysis (DA) Results

Before we carry out classical *quadratic discriminant analysis* (*QDA*), we used the parallel coordinate plot popularized by Wegman [51]. In this method, the similarity lies in that each observation is mapped onto a line. Therefore, the parallel coordinate plots show connected line segments representing each row of a dataset. The x-axis is *p* points - one for each dimension in the data; the y-axis spans from $\min(X)$ to $\max(X)$. What we are doing here is replacing an orthogonal coordinate system with a parallel coordinate system; hence the name.

Looking at the parallel coordinate plot for this dataset, we see that this data set is highly overlapped and also the variables are highly correlated. It is difficult to separate the groups by simply using visual inspection.

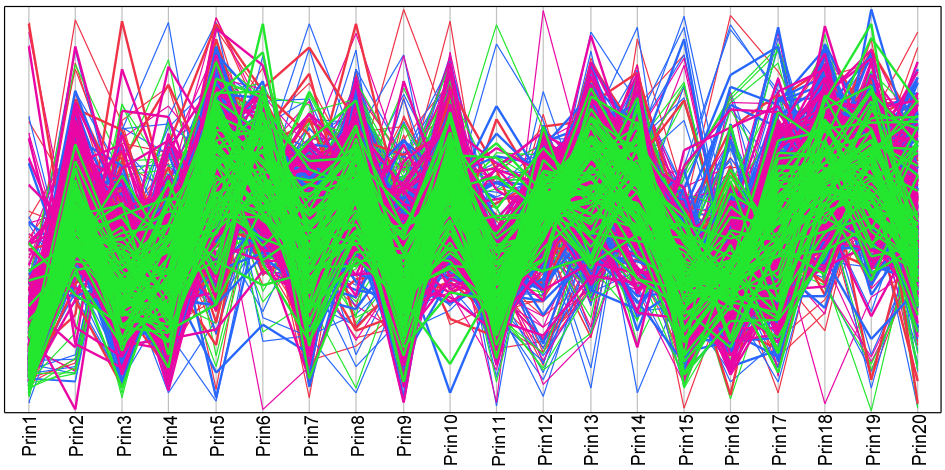

Figure 4.1 Parallel coordinateplot of K=4 mobile phone customer groups.

Next, we used both the *linear discriminant analysis* (*LDA*) and *quadratic discriminant analysis* (*QDA*). These two methods are popular and in general show good performance when data are multivariate normally (or Gaussian) distributed. In our case, since the results from *LDA* and *QDA* were almost the same, and since *QDA* is more general approach where covariance matrices of each group are considered to be different, we only give our results for *QDA*.

Based on the Gaussian assumption, we calculate the posterior probability of group membership of each observation, and assign an observation to a group where the posterior probability of group membership is the greatest. As a result, *QDA* works well in heteroskedastic cases.

Using the maximum probability rule, an observation vector x can be assigned to group *k rather* than *l*, if

$$Q_k\left(x_i\right) > Q_l(x_i) \tag{75}$$

for all $k \neq l$, where

$$Q_k(x_i) = -\frac{1}{2}\left(x_i - \overline{x}_k\right)' S_k^{-1}\left(x_i - \overline{x}_k\right) - \frac{1}{2} log\left|S_k\right| + \log\left(\pi_k\right). \tag{76}$$

Classification result from using *quadratic discriminant analysis* (*QDA*) is summarized in Table 4.1.

Table 4.1 Confusion matrix from Quadratic DA.

Actual\Predicted	$\hat{G}_1$	$\hat{G}_2$	$\hat{G}_3$	$\hat{G}_4$	**Row Total**
G_1	67	**55**	**42**	**43**	207
G_2	**41**	169	**75**	**115**	400
G_3	**35**	**48**	114	**53**	250
G_4	**14**	**31**	**15**	104	164
Column Total	157	303	246	315	$n = 1{,}021$

Looking at Table 4.1, we note that the number of misclassified customers is equal to 567. In other words, 55.53% of the customers of the mobile phone users are misclassified. This is a high number of misclassification rate. The reason for this high misclassification rate is that, the mobile phone dataset is highly non-separable as seen in 2-D and 3-D canonical plots shown in Figures 4.2 and 4.3, respectively.

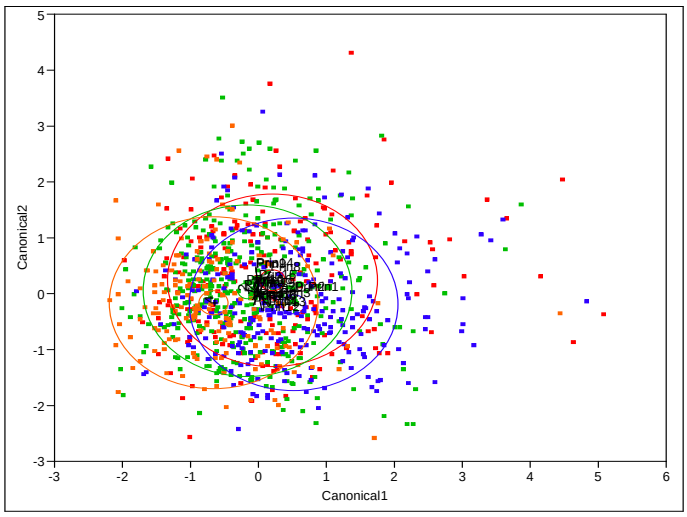

Figure 4.2 2-D Canonical plot of K=4 mobile phone customer groups.

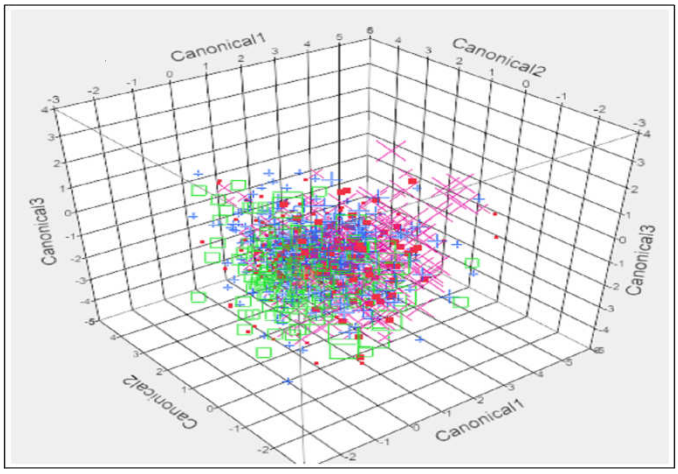

Figure 4.3 3-D Canonical plot of K=4 mobile phone customer groups.

4.4. The Results of the Analysis Using MSVM-RFE

We now use *multi-class* support vector-*recursive feature elimination* (*MSVM-RFE*). The procedure of *MSVM-RFE* is as follows:

- *Train the classifier,*
- *Compute the ranking of ICOMP criterion for features,* and
- *Choose the minimum of ICOMP to select the best subset of the ranked features.*

We use several kernel functions to fit *MSVM-RFE*, such as quadratic, cubic, inverse multi-quadratic, and Cauchy kernels. For this, see Baek and Bozdogan [52]. Table 4.2 shows the results from fitting *MSVM-RFE* using the cubic polynomial kernel for the classification of the mobile phone customer data base and dimension reduction.

Table 4.2 MSVM-RFE Cubic polynomial kernel results.

Methods	*ICOMP*	Train Error	*Test Error*	CI for Train Error	*CI for Test Error*
OAA	138,152	0	**0.724**	[0, 0.655]	[0.621, 0.756]
Subset	2, 5, 14, 15, 17-20 (8)				
OAO	**137,912**	**0**	**0.39**	**[0, 0.197]**	**[0.266, 0.41]**
Subset	2,7,9,10,12,15,18 (7)				
DAG	138,330	0	**0.747**	[0, 0.526]	[0.647, 0.8]
Subset	2, 7,8,9,10,12,13 (7)				
Pairwise	138,606	0	**0.737**	[0, 0.705]	[0.679, 0.794]
Subset	5,7,9,10,13,18,19 (7)				

Note that the test error is 39% using the *cubic polynomial kernel* and *One-Against-One* (*OAO*) *MSVM-RFE* with the best subset of PCs: 2, 7, 9, 10, 12, 15, 18. This is a reduction of dimensionality out of 20 PCs. The other kernels do not give us any better results. *DAG* here refers to *"Directed Acyclic Graph"* and *Pairwise* refers to Pairwise coupling, which is a popular multi-class classification method that combines all comparisons for each pair of classes.

Next, we use the Cauchy kernel as shown in Table 4.3. Note that, *MSVM-RFE* performance the best under the Cauchy kernel. By best, here we mean that both misclassification error rate and also the minimum value of $ICOMP_{PERF}$ are both achieved at the Cauchy kernel.

Table 4.3 MSVM-RFE Cauchy kernel results. Best Solution.

Methods	*ICOMP*	Train Error	*Test Error*	CI for Train Error	*CI for Test Error*
OAA	-138,032	0	**0.604**	[0, 0.377]	[0.56, 0.676]
Subset	1, 3, 5, 7, 8,9,10,13,14 (9)				
OAO	**-137,894**	**0**	**0.227**	**[0, 0.135]**	**[0.165, 0.314]**
Subset	**1, 2, 5, 7, 8,10,13,17,18** (9)				
DAG	-137895	0	**0.578**	[0, 0.447]	[0.559, 0.682]
Subset	1, 2, 4, 7,8,9,10,13,17,18 (10)				
Pairwise	-138034	0	**0.753**	[0, 0.554]	[0.697, 0.803]
Subset	1, 3, 5, 7-10, 16-19 (11)				

Looking at Table 4.3, we see that when we use the *Cauchy kernel* which is the best kernel chosen by *ICOMP* gives us 22.7% misclassification error rate with the best subset of PCs: 1, 2, 5, 7, 8, 10, 13, 17, 18. Although the number PCs chosen increased a bit in fitting the Cauchy kernel as compared to the cubic polynomial kernel, the error rate of classification is improved 16%.

Further we note that when we compare our result with that of the classical *Quadratic Discriminant Analysis* (*QDA*), we have

$$55.53\% - 22.70\% = 32.83\%$$

better results of classification in terms of percent missclassification error rate.

This is a remarkable achievement due to using *MSVM-RFE* hybridized with $ICOMP_{PERF}$ that was not possible before using other methods to classify the mobile phone customer data base as a new micro-marketing analytics.

One now can build a more reliable market segmentation model for mobile phone customer management. Our approach helps mobile phone companies and service providers to sift through their data basis for meaningful relationships by determining the patterns of customer preferences intelligently rather than reacting haphazardly.

5. Conclusions and Discussion

In this paper, we presented several forms of the information-theoretic measure of complexity $ICOMP$ class of criteria. These criteria are based on sound theoretical and technical underpinnings of entropic covariance complexity measure. In this sense, the logical foundations of the procedure presented here are both natural and rational. $ICOMP$ class of criteria refine the original derivation of Akaike's AIC and $AIC-type$ criteria, where Akaike went to the asymptotic distribution of the parameter vector $\hat{\theta}_M$ of the model M too quickly. There are other forms of $ICOMP$ that are robust and at the same time misspecification resistant. For space considerations, we did not show these forms in this paper. Although there is plethora of other model selection criteria in the literature, $ICOMP$ theory is uniquely situated in that it bridges both Frequentist and Bayesian approaches to model selection. One of the many advantages of $ICOMP$ class of criteria is that the use of information-theoretic measure of complexity of the estimated *inverse-Fisher information matrix* $(IFIM)$ avoids the complicated sampling distributions of many well-known classical test statistics, or any table look up, which is potentially of great value in evaluating the goodness of fit of competing models. Therefore, the utility of such model selection criteria is expected to enhance the scientific community. We note that, the development of the $ICOMP$ class of criteria will lead to the acceptance of simpler and scalable, more generalizable, and more precisely estimated models which will have a positive impact on theory development and testing in all fields which utilize such models in their research efforts.

In conclusion, we note that our numerical result on the mobile phone customers data base clearly demonstrate the excellent classification performance of $ICOMP$ over the classical discriminant analysis method using a novel *multi-class support vector machine-recursive feature elimination* (*MSVM-RFE*) method. There are many other applications of $ICOMP$ class of criteria. These are shown and illustrated on many well-known real benchmark and simulated data sets in the forthcoming book of Bozdogan [6] with an accompanying computational toolbox in Matlab.

Acknowledgements

This research was partially supported by the *Jefferson Faculty Prize Award* at the University of Tennessee. The author extends his gratitude and thanks to Prof. Dr. Eyüp Çetin, the Editor-in-Chief of the, *Istanbul University Journal of the School of Business Administration*, for inviting me to make a contribution to this Special Volume of the journal. I acknowledge the help of my doctoral student Mr. Kang Bok Lee, who was able

to type some parts of this paper and fix the formatting issues. Without his help, this paper would not have been completed.

References

[1] H. Bozdogan, ICOMP: A new model-selection criterion. In classification and related methods of data analysis, H. H. Bock (Ed.), Elsevier Science Publishers, Amsterdam, 1988, pp.599-608.

[2] H. Bozdogan, 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 (1990).

[3] H. Bozdogan, Mixture-model cluster analysis using a new informational complexity and model selection criteria. In Multivariate Statistical Modeling, H. Bozdogan (Ed.), Vol. 2, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach, Kluwer Academic Publishers, the Netherlands, Dordrecht, 1994, pp.69-113.

[4] H. Bozdogan, Akaike's information criterion and recent developments in information complexity. Journal of Mathematical Psychology. 44, 62-91 (2000).

[5] H. Bozdogan, Statistical Modeling and Model Evaluation: A New Informational Approach. To appear (2004).

[6] H. Bozdogan, Information Complexity and Multivariate Learning in High Dimensions in Data Mining. To appear (2011).

[7] H. Akaike, Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csáki (Eds.), Second international symposium on information theory, AcadémiaiKiadó, Budapest, 267-281 (1973).

[8] M.H. van Emden, An Analysis of Complexity. Mathematical Centre Tracts, Amsterdam, 35 (1971).

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

[10] H. Bozdogan, Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika. 52, 3, 345-370 (1987).

[11] H. Cramér, Mathematical Methods of Statistics. Princeton University Press, Princeton, NJ, 1946.

[12] C.R. Rao, Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math Soc. 37, 81 (1945).

[13] C.R. Rao, Minimum variance and the estimation of several parameters. Proc. Cam. Phil. Soc. 43, 280 (1947).

[14] C.R. Rao, Sufficient statistics and minimum variance estimates. Proc. Cam. Phil. Soc. 45, 213 (1948).

[15] H. Bozdogan and D.M.A. Haughton, Informational complexity criteria for regression models. Computational Statistics and Data Analysis. 28, 51-76 (1998).

[16] H. Bozdogan and M. Ueno, A unified approach to information-theoretic and Bayesian model selection criteria. Invited paper presented in the Technical Session Track C on: Information Theoretic Methods and Bayesian Modeling at the 6th World

Meeting of the International Society for Bayesian Analysis (ISBA), May 28-June 1, 2000, Hersonissos-Heraklion, Crete (2000).

[17] H. Bozdogan and P. M. Bearse, Subset selection in vector autoregressive models using the genetic algorithm with informational complexity as the fitness function. Systems Analysis, Modeling, Simulation (SAMS) (1998).

[18] S. Kullback, Information Theory and Statistics. Dover, New York, 1968.

[19] C.J. Harris, An information theoretic approach to estimation. In M. J. Gregson (Ed.), Recent Theoretical Developments in Control, Academic Press, London, 1978, pp.563-590.

[20] H. Theil and D.G. Fiebig, Exploiting Continuity: Maximum Entropy Estimation of Continuous Distributions. Ballinger Publishing Company, Cambridge, MA, (1984).

[21] S. Kullback, and R. Leibler, On information and sufficiency. Ann. Math. Statist. 22, 79-86 (1951).

[22] C.E. Shannon, A mathematical theory of communication. Bell Systems Technology Journal, 27, 1948, pp. 379-423.

[23] S. Watanabe, Pattern Recognition: Human and Mechanical. John Wiley and Sons, New York, 1985.

[24] J. Rissanen, Stochastic Complexity in Statistical Inquiry. World Scientific Publishing Company, Teaneck, NJ, 1989.

[25] R.E. Blahut, Principles and Practice of Information Theory. Addison-Wesley Publishing Company, Reading, MA, 1987.

[26] C.R. Rao, Linear Statistical Inference and Its Applications. John Wiley& Sons, New York, 1965, p. 532.

[27] S.A. Mulaik, Linear causal modeling with structural equations, CRC Press, A chapman and Hall Book, 2009, p. 368.

[28] S.A. Mustonen, measure of total variability in multivariate normal distribution. Comp. Statist. and Data Ana. 23, 321-334 (1997).

[29] S.D. Morgera, Information theoretic covariance complexity and its relation to pattern recognition. IEEE Trans. on Syst., Man, and Cybernetics. SMC 15, 608-619 (1985).

[30] J.B. Conway, Functions of one complex variable I, Second edition, Springer-Verlag, 1995.

[31] L. Ljung and J. Rissanen, On canonical forms, parameter identifiability and the concept of complexity. In Identification and System Parameter Estimation, N. S. Rajbman (Ed.), North-Holland, Amsterdam, 1415-1426 (1978).

[32] M. S. Maklad and T.Nichols, A new approach to model structure discrimination. IEEE Trans. on Syst., Man, and Cybernetics. SMC 10, 78-84 (1980).

[33] D.S. Poskitt, Precision, Complexity and Bayesian model determination. J. Roy. Statist. Soc. 49, 199-208 (1987).

[34] B.R. Frieden, Physics from fisher information, Cambridge University press, 1998.

[35] J.Rissanen, Modeling by shortest data description. Automatica, 14, 465-471 (1978).

[36] G.Schwarz, Estimating the dimension of a model. Ann. Statist., 6, 461-464 (1978).

[37] A.D.R. McQuarie, and C-L. Tsai, Regression and Time Series Model Selection. World Scientific Publishing Company, Singapore, 1998.

[38] K.P. Burnham and D. R. Anderson, Model Selection and Inference: A Practical Information-Theoretic Approach. Springer, New York, 1998.

[39] D.V. Lindley, On a measure of information provided by an experiment, The Annals of Mathematical Statistics 27, 4, 986-1005 (1956).

[40] K. Chaloner and I. Verdinelli, Bayesian experimental design a review. Statistical Science. 10, 3, 273-304 (1995).

[41] R.E. Kass, L. Tierney, and J.B. Kadane, The validity of posterior expansions based on Laplace's method. In: GEISSER, S. et al. (Ed.), Bayesian and likelihood methods in statistics and econometrics: essays in honor of George A. Barnard. Amsterdam: North-Holland, 1990. 473-488, 1990.

[42] X. Chen, Model Selection in Nonlinear Regression Analysis. Unpublished Ph.D. Thesis, the University of Tennessee, Knoxville, TN, 1996.

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

[44] J.R.M. Hosking, Language-multiplier tests of time-series models. Journal of the Royal Statistics Society. Series B, 42, 170-181 (1980).

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

[46] A. Howe and H. Bozdogan, Regularized SVM classification with information complexity and the genetic algorithm, appear in multivariate high dimensional data mining forthcoming edited book, 2011.

[47] V. Vapnik, The nature of statistical learning theory, springer-verlag, New York, 1995.

[48] C. Hsu and C. Lin, A comparison of method for multiclass support vector machines. IEEE Transactions on Neural Networks. 13, 2 (2002).

[49] F. Camillo, Personal correspondence, 2007.

[50] F. Camillo, C. Liberati and K.A. Athappilly, Profiling of customer data base through a sample survey, unpublished report, 2009.

[51] E. Wegman, hyperdimensional data analysis using parallel coordinates. Technical Report No.1. George Mason University Center for Computational Statistics (1986).

[52] S.H. Baek and H. Bozdogan, Multi-class support vector machine recursive feature elimination using information complexity, working paper (2011).