This paper develops a computationally feasible intelligent data mining and knowledge discovery technique that addresses the potentially daunting statistical and combinatorial problems presented by subset regression models. Our approach integrates novel statistical modelling procedures based on an information-theoretic measure of complexity. We form a three-way hybrid between: information measures of complexity, multiple regression models, and genetic algorithms (GAs). We demonstrate our new approach using a simulated example and on a real data set to illustrate the versatility and the utility of the new approach.

2.1 Introduction

In regression type problems whether it is in multiple regression analysis, in logistic, or in ordinal logistic regression, model building and evaluation and selection of relevant subset of predictor variables on which to base inferences is a central problem in data mining to reduce the “curse of dimensionality,” a term coined by Richard Bellman (see, Bellman, 1961) almost 42 years ago. Also, see, e.g., Sakamoto et al., 1986, Miller 1990, Boyce et al., 1974. Often a quantitative, binary, or ordinal level
response variable is studied given a set of predictor variables. In such cases it is often desirable to determine which subsets of the predictors are most useful for forecasting the response variable, and to interpret a large number of regression coefficients, since this can become unwieldy even for moderately sized data, and to achieve parsimony of unknown parameters, allowing both better estimation and clearer interpretation of the parameters included in these models.

The problem of selecting the best regression models is a non-trivial exercise, particularly when a large number of predictor variables exist and the researcher lacks precise information about the exact relationships among the variables.

In many cases the total possible number of models reaches over millions (e.g., more than 20 predictor variables) or perhaps into the billions (e.g., 30 predictor variables) and evaluation of all possible combinations of subsets is unrealistic in terms of time and cost.

Therefore, numerical optimization techniques and strategies for model selection are needed to explore the vast solution space. In general the problem of subset selection using numerical techniques requires two components:

1. an algorithm for the efficient searching of the solution space, and
2. a criterion or measure for the comparison of competing models to help guide the search.

Most statistical packages for statistical analysis provide a Backward and Forward stepwise selection strategy for choosing the best subset model. However, it is well known that both Backward and Forward stepwise selection in regression analysis do not always find the best subset of predictor variables from the set of \( k \) variables. Major criticisms levied on Backward and Forward stepwise selection are that, little or no theoretical justification exists for the order in which variables enter or exit the algorithm (Boyce et al., 1974, p. 19, Wilkinson, 1989, p. 177-178), and the arbitrary choices of the probabilities specified a priori to enter and remove the variables in the analysis. Another criticism is that stepwise searching rarely finds the overall best model or even the best subset of a particular size (Mantel, 1970, Hocking, 1976, 1983, Moses, 1986).

Lastly, and most importantly, because only local searching is employed, stepwise selection provides extremely limited sampling from a small area of the vast solution space. Stepwise selection, at the very best, can only produce an “adequate” model (Sokal and Rohlf, 1981, p. 668).

Based on the above shortcomings of existing problems in regression analysis, the purpose of this paper is to introduce and develop a computationally feasible intelligent data mining and knowledge discovery technique based on the genetic algorithm (GA) and information-based model selection criteria for subset selection in multiple regression models. Our approach has been also extended to logistic regression and ordinal logistic regression models as a three-way hybrid. For space considerations, we will report and publish the results of these elsewhere. However, for more on subset selection of best predictors in ordinal logistic regression models, we refer the reader to Lanning and Bozdogan (2003) in this volume.
A GA is a stochastic search algorithm which is based on concepts of biological evolution and natural selection that can be applied to solving problems where vast numbers of possible solutions exist. GAs have been used in a wide variety of fields such as engineering, economics, game theory (Holland, 1992), computational sciences (Forrest, 1993), marketing (Bauer, 1994) and biology (Sumida et al., 1990). Unlike conventional optimization approaches, the GA requires no calculation of the gradient of the objective function and is not likely to be restricted to a local optima (Goldberg, 1989). A GA treats information as a series of codes on a binary string, where each string represents a different solution to a given problem. These strings are analogous to the genetic information coded by genes on a chromosome. A string can be evaluated, according to some “fitness” value, for its particular ability to solve the problem. On the basis of the fitness values strings are either retained or removed from the analysis after each run so that, after many runs, the best solutions have been identified. One important difficulty with any GA is in choosing an appropriate fitness function as the basis for evaluating each solution.

With respect to multiple regression analysis, the fitness value is a subset selection criterion for comparing subset models in a search of the best subset. This can be easily determined by using informational model selection criteria.

The format of this paper is as follows. In Section ??, we discuss what information complexity is, and present its general form in model selection. In Section ??, we give the derived closed form analytical expressions of information complexity $ICOMP$ of Bozdogan (1988, 1990, 1994, 2000), and Akaike’s (1973, 1974) information criterion $AIC$, and Rissanen’s (1978, 1986) $MDL$, Schwarz’s (1978) $SBC$ or $BIC$, and Bozdogan’s (1987) $Consistent AIC with Fisher information CAICF$ as decision rules for model selection and evaluation in multiple regression models. In Section ??, we develop the GA for the general regression modelling and discuss the new statistical software we developed with graphical user interface (GUI) in a flexible Matlab computational environment. Section ?? is devoted to simulated and real data examples in multiple regression models to demonstrate the versatility and utility of our new approach. In Section ??, we draw our conclusions.

### 2.2 What is Information Complexity: ICOMP?

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. In this section, we develop and present information-
theoretic ideas of a measure of “overall” model complexity in statistical modelling to help provide new approaches relevant to statistical inference.

Recently, based on Akaike’s (1973) original AIC, many model-selection procedures that take the form of a penalized likelihood (a negative log likelihood plus a penalty term) have been proposed (Sclove, 1987). For example, for AIC this form is given by

\[ AIC(k) = -2 \log L(\hat{\theta}_k) + 2m(k) \]  

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

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 are used. In using AIC, according to Akaike (1987, p. 319), the accuracy of parameter estimates is measured by a universal criterion, namely

\[ \text{Accuracy Measure} = E \left[ \log \text{likelihood of the fitted model} \right] \]  

where \( E \) denotes the expectation, since AIC is an unbiased estimator of minus twice the expected log likelihood.

We are motivated from considerations similar to those in AIC. However, 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

\[ \text{Statistical model} = \text{Signal} + \text{Noise} \]  

ICOMP is designed to estimate a loss function:

\[ \text{Loss} = \text{Lack of Fit} + \text{Lack of Parsimony} + \text{Profusion of Complexity} \]  

in several ways using the additivity properties of information theory. We further base our developments on similar considerations to 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 (1987).

The development and construction of ICOMP is based on a generalization of the covariance complexity index originally introduced by van Emden (1971). Instead of penalizing the number of free parameters directly, ICOMP penalizes the covariance complexity of the model. It is defined by

\[ ICOMP = -2 \log L(\hat{\theta}) + 2C(\hat{\Sigma}_{\text{Model}}) \]
Statistical Data Mining and Knowledge Discovery

where \( L(\hat{\theta}_k) \) is the maximized likelihood function, \( \hat{\theta}_k \) is the maximum likelihood estimate of the parameter vector \( \theta_k \) under the model \( M_k \), and \( C \) represents a real-valued complexity measure and \( \hat{\text{Cov}}(\hat{\theta}) = \hat{\Sigma}_{\text{Model}} \) represents the estimated covariance matrix of the parameter vector of the model.

Since there are several forms and justifications of ICOMP, based on (\ref{eq:ICOMP_def}), in this paper, for brevity, we will present the most general form of ICOMP referred to as ICOMP(IFIM). ICOMP(IFIM) exploits the well-known asymptotic optimality properties of the MLE’s, and uses the information-based complexity of the inverse-Fisher information matrix (IFIM) of a model. This is known as the celebrated Cramér-Rao lower bound (CRLB) matrix. See, e.g., Cramér (1946) and Rao (1945, 1947, 1948).

Before we derive ICOMP(IFIM), we first introduce some background material to understand the concept of complexity and give the definition of the complexity of a system next.

2.2.1 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.

An understanding of complexity is necessary in general model building theory and inductive inference to study uncertainty in light of the data. Statistical models and methods are not exactly deductive since human beings often reason on the basis of uncertainties. Instead, they generally fall under the category of inductive inference. Inductive inference is the problem of choosing a parameter, or model, from a hypothesis, or model space, which best ‘explains’ the data under study (Baxter, 1996, p. 1). As discussed in Akaike (1994, p. 27), reasoning under uncertainty was studied by the philosopher C. S. Pierce (see, e.g., Pierce, 1955), who called it the logic of abduction, or in short, abduction. Abduction is a way of reasoning that uses general principles and observed facts to obtain new facts, but all with a degree of uncertainty. Abduction takes place using numerical functions and measures such as the information theoretic model selection criteria. Pierce insisted that the most original part of scientific work was related to the abductive phase, or the phase of selection of proper hypotheses. Therefore, developing a systematic procedure for
abductive inference with the aid of the notion of complexity is “a prerequisite to the understanding of learning and evolutionary processes” (von Neumann, 1966). In this context, statistical modelling and model building is a science of abduction which forms the philosophical foundation of data mining and knowledge discovery. Hence, the study of complexity is of considerable practical importance for model selection of proper hypotheses or models within the data mining enterprise.

We give the following simple system theoretic definition of complexity to motivate a statistically defined measure.

Definition 2.1. 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 frequently now 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 complexity is defined in Definition 2.1, 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, p. 7 and 8), and Bozdogan (1990).

2.2.2 Information Theoretic Measure of Complexity of a Multivariate Distribution

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

Definition 2.2. 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 \( f(x) = f(x_1, x_2, \ldots, x_p) \) and marginal density functions \( f_j(x_j), j = 1, 2, \ldots, p \). Following Kullback (1968), Harris (1978), Theil and Fiebig (1984), and others, we define the informational measure of dependence between random variables \( x_1, x_2, \ldots, x_p \) by
\[
I(x) = I(x_1, x_2, \ldots, x_p) = E_f \left[ \log \frac{f(x_1, x_2, \ldots, x_p)}{f_1(x_1)f_2(x_2)\cdots f_p(x_p)} \right].
\]

Or, it is equivalently defined by
\[
I(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_p) \log \frac{f(x_1, x_2, \ldots, x_p)}{f_1(x_1)f_2(x_2)\cdots f_p(x_p)} dx_1 \cdots dx_p, \quad (2.6)
\]
where \( I \) is the Kullback-Leibler (KL) (1951) information divergence against independence. \( I(x) \) in (2.6) 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.** \( I(x) \equiv I(x_1, x_2, \ldots, x_p) \geq 0 \), i.e., the expected mutual information is nonnegative.

- **Property 2.** \( f(x_1, x_2, \ldots, x_p) = f_1(x_1)f_2(x_2)\cdots f_p(x_p) \) for every p-tuple \( (x_1, x_2, \ldots, x_p) \) if and only if the random variables \( x_1, x_2, \ldots, x_p \) are mutually statistically independent. In this case the quotient in (2.6) is equal to unity, and its logarithm is then zero. Hence, \( I(x) \equiv I(x_1, x_2, \ldots, x_p) = 0 \). If it is not zero, this implies a dependency.

We relate the KL divergence in (2.6) to Shannon’s (1948) entropy by the important identity
\[
I(x) \equiv I(x_1, x_2, \ldots, x_p) = \sum_{j=1}^{p} H(x_j) - H(x_1, x_2, \ldots, x_p), \quad (2.7)
\]
where \( H(x_j) \) is the marginal entropy, and \( H(x_1, x_2, \ldots, x_p) \) is the global or joint entropy. Watanabe (1985) calls (2.7) the strength of structure and a measure of interdependence. We note that (2.7) is the sum of the interactions in a system with \( x_1, x_2, \ldots, x_p \) as components, which we define to be the entropy complexity of that system. This is also called the Shannon Complexity (see, Rissanen, 1989). The more interdependency in the structure, the larger will be the sum of the marginal entropies to 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).

To define the information-theoretic measure of complexity of a multivariate distribution, we let \( f(x) \) be a multivariate normal density function given by
where $\mu = (\mu_1, \mu_2, \ldots, \mu_p)'$, $-\infty < \mu_j < \infty$, $j = 1, 2, \ldots, p$ and $\Sigma > 0$ (p.d.). We write $x \sim N_p(\mu, \Sigma)$. Then the joint entropy $H(x) = H(x_1, x_2, \ldots, x_p)$ from (2.8) for the case in which $\mu = 0$ is given by

$$H(x) = H(x_1, x_2, \ldots, x_p) = - \int f(x) \log f(x) \, dx$$

$$= \int f(x) \left[ \frac{p}{2} \log(2\pi) + \frac{1}{2} \frac{1}{|\Sigma|} \right] \, dx$$

$$= \frac{p}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma|.$$  

(2.9)  

Then, since $E[(x - \mu)(x - \mu)'] = \Sigma$, we have

$$H(x) = H(x_1, x_2, \ldots, x_p) = \frac{p}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma| + \frac{1}{2} \log(\sigma_j^2), j = 1, 2, \ldots, p.$$  

(2.10)  

2.2.3 Initial Definition of Covariance Complexity

Van Emden (1971, p.61) provides a reasonable initial definition of informational complexity of a covariance matrix $\Sigma$ for the multivariate normal distribution. This measure is given by:

$$I(x_1, x_2, \ldots, x_p) = \sum_{j=1}^{p} H(x_j) - H(x_1, x_2, \ldots, x_p)$$

$$= \sum_{j=1}^{p} \left[ \frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\sigma_j^2) + \frac{1}{2} \right] - \frac{p}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{p}{2}.$$  

(2.12)
This reduces to

\[
C_0(\Sigma) = \frac{1}{2} \sum_{j=1}^{p} \log(\sigma_{jj}) - \frac{1}{2} \log |\Sigma|,
\]

where \(\sigma_{jj} \equiv \sigma_j^2\) is the \(j\)-th diagonal element of \(\Sigma\) and \(p\) is the dimension of \(\Sigma\). Note that \(C_0(\Sigma) = 0\) when \(\Sigma\) 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 \(\theta = (\theta_1, \theta_2, \ldots, \theta_k)\) 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 \(\theta\) and the product of the marginal densities of the components of \(\theta\). As pointed out by van Emden (1971), the result in (2.13) is not an effective measure of the amount of complexity in the covariance matrix \(\Sigma\), since:

- \(C_0(\Sigma)\) depends on the marginal and common distributions of the random variables \(x_1, \ldots, x_p\), and
- The first term of \(C_0(\Sigma)\) in (2.13) would change under orthonormal transformations.

### 2.2.4 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 \(\Sigma\) should be independent of the coordinates of the original random variables \((x_1, x_2, \ldots, x_p)\) associated with the variances \(\sigma_j^2\), \(j = 1, 2, \ldots, p\). As it is \(C_0(\Sigma)\) in (2.13) is coordinate dependent. However, to characterize the maximal amount of complexity of \(\Sigma\), we can relate the general definition of complexity of \(\Sigma\) to the total amount of interaction or \(C_0(\Sigma)\) in (2.13). We do this by recognizing the fact that the maximum of (2.13) under orthonormal transformations of the coordinate system may reasonably serve as the measure of complexity of \(\Sigma\). This corresponds to observing the interaction between the variables under the coordinate system that most clearly represents it in terms of the measure \(I(x_1, x_2, \ldots, x_p) \equiv C_0(\Sigma)\). So, to improve on (2.13), we have the following proposition.

**Proposition 2.1.** A maximal information theoretic measure of complexity of a covariance matrix \(\Sigma\) of a multivariate normal distribution is

\[
C_1(\Sigma) = \max_{\mathcal{T}} C_0(\Sigma) = \max_{\mathcal{T}} \left\{ H(x_1) + \ldots + H(x_p) - H(x_1, x_2, \ldots, x_p) \right\}
= \frac{1}{2} \log \left[ \frac{\text{tr}(\Sigma)}{p} \right] - \frac{1}{2} \log |\Sigma|,
\]

(2.14)
where the maximum is taken over the orthonormal transformation $T$ of the overall coordinate systems $x_1, x_2, \ldots, x_p$.

**Proof:** Following van Emden (1971, p. 61), Ljung and Rissanen (1978, p. 1421), and filling the gap in Maklad and Nichols (1980, p. 82) to find

$$C_1(\Sigma) = \max_T \{H(x_1) + \ldots + H(x_p) - H(x_1, x_2, \ldots, x_p)\} \quad (2.15)$$

we must find the orthonormal transformation, say $T$, of $\Sigma$ that maximizes

$$\sum_{j=1}^{p} \log(\sigma_{jj}^*) + \ldots + \log(\sigma_{pp}^*), \quad (2.16)$$

where $\sigma_{jj}^* = \sigma_j^{2*}$'s are the diagonal elements of the covariance of $Tx = T(x_1, x_2, \ldots, x_p)$, i.e., $\text{Cov}(Tx) = \Sigma^*$. Since orthonormal transformations leave $tr(\Sigma) = \sigma_{11} + \ldots + \sigma_{pp}$ invariant, we

$$\text{maximize } \sum_{j=1}^{p} \log(\sigma_{jj}) \quad (2.17)$$

**subject to** $tr(\Sigma) = c, c = \text{constant}$. 

To carry out this maximization, we use the geometric and arithmetic mean of $\sigma_{11}, \ldots, \sigma_{pp}$ given by

$$\left(\prod_{j=1}^{p} \sigma_{jj}\right)^{\frac{1}{p}} \leq \frac{1}{p} \sum_{j=1}^{p} \sigma_{jj} \quad (2.18)$$

with equality if and only if $\sigma_{11} = \sigma_{22} = \ldots = \sigma_{pp}$.

The equality condition in (2.17) is always achieved by orthonormal transformations $T$ to equalize all variances to within certain error. This is shown by van Emden (1971, p.66). Hence, from van Emden (1971, p.61), and Maklad and Nichols (1980, p.82), (2.17) implies that

$$\max_{j=1}^{p} \sum_{j=1}^{p} \log(\sigma_{jj}) = \max_{j=1}^{p} \log \left(\prod_{j=1}^{p} \sigma_{jj}\right)$$

$$= p \log tr(\Sigma) - p \log p = p \log \left(\frac{tr(\Sigma)}{p}\right). \quad (2.19)$$

Now replacing the first component of $C_0(\Sigma)$ in (2.12), we find
\[ C_1(\Sigma) = \max_T \{ H(x_1) + \ldots + H(x_p) - H(x_1, x_2, \ldots, x_p) \} \]

\[ = \frac{p}{2} \log \left( \frac{\text{tr}(\Sigma)}{p} \right) - \frac{1}{2} \log |\Sigma| \tag{2.20} \]

as a maximal information theoretic measure of the complexity of a covariance matrix \( \Sigma \) of a multivariate normal distribution.

\( C_1(\Sigma) \) in (2.20) is an upper bound to \( C_0(\Sigma) \) in (2.19), and it measures both inequality among the variances and the contribution of the covariances in \( \Sigma \) (van Emden, 1971, p.63). Such a measure is very important in model selection and evaluation problems 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 \( \sigma_j^2 \equiv \sigma_{jj}, j = 1, 2, \ldots, p \). Furthermore, if, for example, one of the \( \sigma_j^2 \)'s is equal to zero, then \( C_0(\Sigma) \) in (2.19) takes the value “\( \infty - \infty \)” which is “indeterminate,” whereas \( C_1(\Sigma) \) in (2.20) has the value “\( \infty \)” (infinity) which has a mathematical meaning. Also, \( C_1(\Sigma) \) in (2.20) 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 \( \Sigma \); see Magnus and Neudecker (1999, p. 26). These properties are given and established in Bozdogan (1990).

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 stable parameter is equivalent to the simplicity of model covariance structure defined in Bozdogan (1990). 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 \( \Sigma \), and that we do not discard the use of \( C_0(\Sigma) \). If we write (2.20) as

\[ C_1(\Sigma) = \frac{1}{2} \log \left( \frac{\text{tr}(\Sigma)^p}{|\Sigma|} \right) \tag{2.21} \]

we interpret the complexity as the \( \log \) ratio between the geometric mean of the average total variation and the generalized variance, since \( \text{tr}(\Sigma)/p \) is equal to average total variation, and \( |\Sigma| \) is the generalized variance.

Further, if we let \( \lambda_1, \lambda_2, \ldots, \lambda_p \) be the eigenvalues of \( \Sigma \), then \( \text{tr}(\Sigma)/p = \overline{\lambda}_a = 1/p \sum_{j=1}^p \lambda_j \) is the arithmetic mean of the eigenvalues of \( \Sigma \), and \( |\Sigma|^{1/p} = \overline{\lambda}_g = \left( \prod_{j=1}^p \lambda_j \right)^{1/p} \) is the geometric mean of the eigenvalues of \( \Sigma \). Then the complexity of \( \Sigma \) can be written as

\[ C_1(\Sigma) = \frac{p}{2} \log \left( \frac{\overline{\lambda}_a}{\overline{\lambda}_g} \right) \tag{2.22} \]
Hence, we interpret the complexity as the log ratio between the arithmetic mean and the geometric mean of the eigenvalues of $\Sigma$. It measures how unequal the eigenvalues of $\Sigma$ 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 $\Sigma$ (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, $C_1(\Sigma) \rightarrow 0$ as $\Sigma \rightarrow I$, the identity matrix. This establishes a plausible relation between information-theoretic complexity and computational effort. Further, 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 collinearity are the least complex, or most informative, and that 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 $\Sigma$. An orthogonal transformation $T$ indeed exists which corresponds to a sequence of plane rotation of the coordinate axes to equalize the variances. This can be achieved using Jacobi’s iterative method or 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 $R$ be the correlation matrix obtained from $\Sigma$ by the relationship

$$R = \Lambda_{\sigma} \Sigma \Lambda_{\sigma},$$  

(2.23)

where $\Lambda_{\sigma} = \text{diag}(1/\sigma_1, \ldots, 1/\sigma_p)$ is a diagonal matrix whose diagonal elements equal to $1/\sigma_j, j = 1, 2, \ldots, p$. From (2.23), we have

$$C_1(R) = -1/2 \log |R| \equiv C_0(R).$$  

(2.24)

Diagonal operation of a covariance matrix $\Sigma$ always reduces the complexity of $\Sigma$, and that $C_1(R) \equiv C_0(R)$ takes into account the interdependencies (correlations) among the variables. For simplicity, the $C_0$ measure based on the correlation matrix $R$ will be denoted by $C_R$, and $C_0(R)$ is written as $C_R(\Sigma)$ for notational convenience, since $R$ is obtained from $\Sigma$. 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 $|R| = 1$, then $I(x_1, x_2, \ldots, x_p) = 0$ which implies the mutual independence of the variables $x_1, x_2, \ldots, x_p$. If the variables are not mutually independent, then $0 < |R| < 1$ and that $I(x_1, x_2, \ldots, x_p) > 0$. In this sense $I(x)$ in (2.20) or (2.21) can also be viewed as a measure of dimensionality of model manifolds.
Next, we develop the informational complexity $ICOMP(IFIM)$ approach to model evaluation based on the maximal covariance complexity $C_1(\bullet)$, and $C_R(\bullet)$.

### 2.2.5 ICOMP as an Approximation to the Sum of Two Kullback-Leibler Distances

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.

**Proposition 2.2.** For a multivariate normal linear or nonlinear structural model we define the general form of $ICOMP(IFIM)$ as

$$ICOMP(IFIM) = -2\log L(\hat{\theta}) + 2C_1(\hat{F}^{-1}(\hat{\theta})), \quad (2.25)$$

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

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

$$y = m(\theta) + \epsilon, \quad (2.26)$$

where:

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

$$E(\epsilon) = 0, E(\epsilon\epsilon') = \Sigma_\epsilon. \quad (2.27)$$

Following Bozdogan and Haughton (1998), we denote $\theta^*$ to be a vector of parameters of the operating true model, and $\theta$ to be any other value of the vector of parameters. Let $f(y; \theta)$ denote the joint density function of $y$ given $\theta$. Let $I(\theta^*; \theta)$ denote the KL distance between the densities $f(y; \theta^*)$ and $f(y; \theta)$. Then, since $y_i$ are independent, $i = 1, 2, \ldots, n$, we have
\[ I(\theta^*; \theta) = \int_{\mathbb{R}^n} f(y; \theta^*) \log \left( \frac{f(y; \theta^*)}{f(y; \theta)} \right) dy \]
\[ = \sum_{i=1}^{n} \int f_i(y_i; \theta^*) \log \left[ f_i(y_i; \theta^*) \right] dy_i - \sum_{i=1}^{n} \int f_i(y_i; \theta) \log \left[ f_i(y_i; \theta) \right] dy_i, \quad (2.28) \]

where \( f_i, i = 1, 2, \ldots, n \) are the marginal densities of the \( y_i \).

Note that the first term in (2.28) 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:
\[ -\sum_{i=1}^{n} E \left[ \log f_i(y_i; \theta) \right], \quad (2.29) \]
which can be unbiasedly estimated by
\[ -\sum_{i=1}^{n} \log f_i(y_i; \theta) = -\log L(\theta | y_i), \quad (2.30) \]
where \( \log L(\theta | y_i) \) is the log likelihood function of the observations evaluated at \( \theta \).

Given a model \( M \) where the parameter vector is restricted, a maximum likelihood estimator \( \hat{\theta}_M \) can be obtained for \( \theta \), and the quantity
\[ -2 \sum_{i=1}^{n} \log f_i(y_i; \hat{\theta}_M) = -2 \log L(\hat{\theta}_M) \quad (2.31) \]
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 \( \text{Cov}(\hat{\theta}_M) = \Sigma(\hat{\theta}_M) \) for the MLE \( \hat{\theta}_M \). That is,
\[ \hat{\theta}_M \sim N\left( \theta^*, \Sigma(\hat{\theta}_M) \equiv \hat{F}^{-1}(\hat{\theta}_M) \right). \quad (2.32) \]

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 (2.28), 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:
\[
ICOMP(\text{IFIM}) = -2 \sum_{i=1}^{n} \log f_i(y_i; \hat{\theta}_M) + 2C_1\left( \hat{\Sigma}(\hat{\theta}_M) \right) \\
= -2 \log L(\hat{\theta}_M) + 2C_1\left( \hat{F}^{-1}(\hat{\theta}_M) \right). \quad (2.33)
\]

The first component of \( ICOMP(\text{IFIM}) \) in (2.33) 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 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. See, e.g., Cramér (1946) and Rao (1945, 1947, 1948).

This approach has several rather attractive features. If $F_{jj}^{-1}(\theta_k)$ is the j-th diagonal element of the inverse-Fisher information matrix (IFIM), from Chernoff (1956), we know that $F_{jj}^{-1}(\theta_j)$ 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

$$F_{jj}^{-1}(\theta_K) \geq F_{jj}^{-1}(\theta_k).$$

Behboodian (1964) explains that the inequality (2.34) 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. This is an important result that impacts the parameter redundancy. The proof of (2.34) is shown in Chen (1996, p. 6) in his doctoral dissertation: “Model Selection in Nonlinear Regression Analysis” under my supervision.

The use of the $C_1(\hat{F}^{-1}(\hat{\theta}_M))$ 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 arithmetical means of the eigenvalues of IFIM given by

$$ICOMP(IFIM) = -2\log L(\hat{\theta}_M) + s\log \left(\frac{\lambda_a}{\lambda_g}\right),$$

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

We note that $ICOMP(IFIM)$ now looks in appearance like the CAIC of Bozdogan (1987), Rissanen’s (1978) MDL, and Schwarz’s (1978) Bayesian criterion SBC, except for using $\log \left(\frac{\lambda_a}{\lambda_g}\right)$ instead of using $\log(n)$, where $\log(n)$ denotes the natural logarithm of the sample size $n$. A model with minimum $ICOMP$ is chosen to be the best among all possible competing alternative models.

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 (??) penalizes the “bad scaling” of the parameters. It is important to note that well conditioning of the information matrix needs a simple structure, but the latter does not necessarily imply the former. For example, consider an information matrix that 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 often ignores an important characteristic, namely, the ratios of the diagonal elements in the information matrix, or the "scale" of these components.

Similar to AIC, to make ICOMP(IFIM) to be scale invariant with respect to scaling and orthonormal transformations in model selection enterprise, we suggest the use of the correlational form of IFIM given by

\[
\hat{F}^{-1}(\hat{\theta}) = D_{F^{-1}F}^{-1/2}D_{F^{-1}}^{-1/2}.
\]

Then, ICOMP(IFIM)_R is defined by

\[
ICOMP(IFIM)_R = -2\log L(\hat{\theta}) + 2C_1(\hat{F}^{-1}(\hat{\theta}_M)) - 2C_1(\hat{F}^{-1}R(\hat{\theta}_M)).
\]

In this way ICOMP becomes invariant to one-to-one transformations of the parameter estimates. In the literature, several authors such as McQuarie and Tsai (1998, p. 367), and Burnham and Anderson (1998, p.69), without reviewing the impact and the applications of ICOMP to many complex modelling problems, have erroneously interpreted the contribution of this novel approach over AIC, and AIC-type criteria.

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 (1987) Consistent AIC (CAIC). The lack of parsimony is automatically adjusted by \(C_1(\hat{F}^{-1}(\hat{\theta}_M))\) or \(C_1(\hat{F}^{-1}R(\hat{\theta}_M))\) across the competing alternative portfolio of models as the parameter spaces of these models are constrained in the model selection process.

Following Morgera (1985, p. 612), we define the relative reduction of complexity (RRC) in terms of the estimated IFIM as

\[
RRC = \frac{C_1(\hat{F}^{-1}(\hat{\theta}_M)) - C_1(\hat{F}^{-1}R(\hat{\theta}_M))}{C_1(\hat{F}^{-1}(\hat{\theta}_M))},
\]

and the percent relative reduction of complexity by

\[
PRRC = \frac{C_1(\hat{F}^{-1}(\hat{\theta}_M)) - C_1(\hat{F}^{-1}R(\hat{\theta}_M))}{C_1(\hat{F}^{-1}(\hat{\theta}_M))} \times 100%.
\]

The interpretation of RRC or PRRC is that they both measure heteroscedastic complexity plus a correlational complexity of the model. In general statistical modelling framework, what this means is that, when the parameter estimates are highly correlated, in nonlinear, and in many other statistical modelling, one can remove the correlation by considering parameter transformations of the model. The difference between the complexities \(C_1(\hat{F}^{-1}(\hat{\theta}_M))\) and \(C_1(\hat{F}^{-1}R(\hat{\theta}_M))\) 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 scaling effect. Indeed, the reduction in the correlation may even make scaling worse. 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 these, see, e.g., Chen (1996), Bozdogan (2000), and Chen and Bozdogan (2004).

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 and Bayesian justification of $ICOMP$. These versions of $ICOMP$ are useful in linear and nonlinear models. For more details on this and other approaches, we refer the readers to Bozdogan (2000), and Bozdogan and Haughton (1998) where consistency properties of $ICOMP$ have been studied in the case of the usual multiple regression models. The probabilities of underfitting and overfitting for $ICOMP$ 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 the 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.

In concluding this section, we note that 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 the unbiased estimates of the parameters. Further, we have the advantage of using smoothed (or improved) covariance estimators of the models and measure the complexities to study the robustness properties of different methods of parameter estimates and improved covariance estimators. $AIC$ 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 legitimize the role of the Fisher information matrix (FIM) (a tribute to Rao, 1945, 1947, 1948) as the natural metric on the parameter manifold of the model, which remained academic in the statistical literature.

2.3 Information Criteria for Multiple Regression Models

We consider the multiple linear regression model in matrix form given by

$$ y = X\beta + \varepsilon $$

(2.40)

where $y$ is a vector of $(n \times 1)$ observations on a dependent variable, $X$ is a full rank $(n \times q)$ matrix of nonstochastic predetermined variables in standardized form, and
\( \beta \) is a \((q \times 1)\) coefficient vector, and \( \varepsilon \) is an \((n \times 1)\) vector of unknown disturbance term, such that

\[ \varepsilon \sim N(0, \sigma^2 I) \text{ or equivalently } \varepsilon_i \sim N(0, \sigma^2) \text{ for } i = 1, 2, \ldots, n. \] (2.41)

Given the model in (2.2) under the assumption of normality, we can analytically express the density function of regression model for a particular sample observation as

\[ f(y_i|x_i, \beta, \sigma^2) = (2\pi \sigma^2)^{-\frac{1}{2}} \exp \left[ -\frac{(y_i - X'\beta)^2}{2\sigma^2} \right] . \] (2.42)

That is, the random observation vector \( y \) is distributed as a multivariate normal with mean vector \( X\beta \) and covariance matrix \( \sigma^2 I_n \). The likelihood function of the sample is:

\[ L(\beta, \sigma^2|y, X) = (2\pi \sigma^2)^{-\frac{n}{2}} \exp \left[ -\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2} \right] , \] (2.43)

and the log likelihood function is:

\[ l(\beta, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{(y - X\beta)'(y - X\beta)}{2\sigma^2} . \] (2.44)

Using matrix differential calculus of Magnus and Neudecker (1999), the maximum likelihood estimates (MLE’s) \( (\hat{\beta}, \hat{\sigma}^2) \) of \( (\beta, \sigma^2) \) are given by:

\[ \hat{\beta} = (X'X)^{-1}X'y, \text{ and } \] (2.45)

\[ \hat{\sigma}^2 = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{n} = \frac{RSS}{n} . \] (2.46)

The maximum likelihood (ML) covariance matrix of the estimated regression coefficients is given by

\[ \hat{Cov}(\hat{\beta})_{MLE} = \hat{\sigma}^2(X'X)^{-1} \] (2.47)

without centering and scaling the model matrix \( X \). Also, the inverse Fisher information matrix (IFIM) is given by

\[ \hat{Cov}(\hat{\beta}, \hat{\sigma}^2) = \hat{F}^{-1} = \begin{bmatrix} \hat{\sigma}^2(X'X)^{-1} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix} . \] (2.48)

Now, we can define derived forms of ICOMP in multiple regression as follows. This can be defined for both \( \hat{Cov}(\hat{\beta}) \), \( \hat{F}^{-1} \), and the correlational form of IFIM, \( \hat{F}_R^{-1} \). These are as follows.
2.3.1 ICOMP Based on Complexity Measures

\[ ICOMP(\text{Reg})_{C_0} = -2 \log L(\hat{\theta}) + 2C_0(\hat{\text{Cov}}(\hat{\beta})) \]

\[ = n \log(2\pi) + n \log(\hat{\sigma}^2) + n + 2\left[ \frac{1}{2} \sum_{j=1}^{q} \log(\hat{\sigma}_{jj}(\hat{\beta})) - \frac{1}{2} \log \left| \hat{\text{Cov}}(\hat{\beta}) \right| \right] \]

\[ = n \log(2\pi) + n \log(\hat{\sigma}^2) + n + 2\left[ \frac{1}{2} \sum_{j=1}^{q} \log(\hat{\sigma}_{jj}(\hat{\beta})) - \frac{1}{2} \sum_{j=1}^{q} \log(\hat{\lambda}_j) \right]. \quad (2.49) \]

based on the original definition of the complexity \( C_0(\cdot) \) in (??).

\[ ICOMP(\text{Reg})_{C_1} = -2 \log L(\hat{\theta}) + 2C_1(\hat{\text{Cov}}(\hat{\beta})) \]

\[ = n \log(2\pi) + n \log(\hat{\sigma}^2) + n + 2\left[ \frac{1}{2} \log \left( \frac{\text{tr}(\hat{\text{Cov}}(\hat{\beta}))}{q} \right) - \frac{1}{2} \log \left| \hat{\text{Cov}}(\hat{\beta}) \right| \right] \]

\[ = n \log(2\pi) + n \log(\hat{\sigma}^2) + n + 2\left[ \frac{1}{2} \log(\hat{\lambda}_q) \right]. \quad (2.50) \]

based on \( C_1(\cdot) \) in (??).

If we use the estimated inverse Fisher information matrix (IFIM) in (??), then we define \( ICOMP(\text{IFIM}) \) as

\[ ICOMP(\text{IFIM})_{\text{Regression}} = -2 \log L(\hat{\theta}_M) + 2C_1\left( \hat{F}^{-1}(\hat{\theta}_M) \right) \]

\[ = n \log(2\pi) + n \log(\hat{\sigma}^2) + n + C_1\left( \hat{F}^{-1}(\hat{\theta}_M) \right). \quad (2.51) \]

where

\[ C_1\left( \hat{F}^{-1}(\hat{\theta}_M) \right) = (q + 1) \log \left[ \frac{\text{tr}\left( \hat{\sigma}^2(X'X)^{-1} + \frac{2\hat{\sigma}^4}{n} \right) \hat{\sigma}^2(X'X)^{-1} \hat{\sigma}^2}{q + \frac{2\hat{\sigma}^4}{n}} \right] \]

\[ - \log(\hat{\sigma}^2(X'X)^{-1}) - \log(\frac{2\hat{\sigma}^4}{n}). \quad (2.52) \]

In (??), as the number of parameters increases (i.e., as the size of \( X \) increases), the error variance \( \hat{\sigma}^2 \) gets smaller even though the complexity gets larger. Also, as
increases, \((X'X)^{-1}\) decreases. Therefore, \(C_1(\hat{F}^{-1})\) achieves a trade-off between these two extremes and guards against multicollinearity.

To preserve scale invariance, we use the correlational form of IFIM, that is, we use \(\hat{F}^{-1}\) and define the correlational form of ICOMP(IFIM)\textsubscript{Regression} given by

\[
ICOMP(IFIM)_{\text{Regression}} = n \log(2\pi) + n \log(\hat{\sigma}^2) + n + C_1(\hat{F}^{-1}(\hat{\theta}_M)) \tag{2.53}
\]

where

\[
\hat{F}^{-1}(\hat{\theta}_M) = D_{F^{-1}}^{-1/2} F^{-1} D_{F^{-1}}^{-1/2} = USV' \tag{2.54}
\]

using the singular value decomposition (svd) on \(\hat{F}^{-1}\). Svd(\() produces a diagonal matrix \(S\), of the same dimension as \(\hat{F}^{-1}\) and with nonnegative diagonal elements in decreasing order, which are the singular values of \(\hat{F}^{-1}\), and unitary matrices \(U\) and \(V\), which satisfy \(UU' = VV' = I\), so that \(\hat{F}^{-1} = U \cdot S \cdot V'\).

### 2.3.2 ICOMP Under Misspecification

Although we will not use this form of ICOMP in this paper, to be complete and to inform the readers, when the model is misspecified, we define ICOMP under misspecification as

\[
ICOMP(IFIM)_{\text{Misspec}} = n \log(2\pi) + n \log(\hat{\sigma}^2) + n + 2C_1(\hat{\text{Cov}}(\hat{\theta}))_{\text{Misspec}}, \tag{2.55}
\]

where

\[
\hat{\text{Cov}}(\hat{\theta})_{\text{Misspec}} = \hat{F}^{-1}\hat{R}\hat{F}^{-1} \tag{2.56}
\]

is a consistent estimator of the covariance matrix \(\text{Cov}(\hat{\theta}_k)\). This is often called the “sandwiched covariance” or “robust covariance” estimator, since it is a correct covariance regardless of whether the assumed model is correct or not. It is called sandwiched covariance, because \(\hat{R}\) is the meat and the two \(\hat{F}^{-1}\)s are slices of the bread. When the model is correct we get \(\hat{F} = \hat{R}\), and the formula reduces to the usual inverse Fisher information matrix \(\hat{F}^{-1}\) (White, 1982).

In the regression case, the Fisher information in inner-product form is given as in (??) by

\[
\hat{F}^{-1} = \begin{bmatrix}
\hat{\sigma}^2 (X'X)^{-1} & 0 \\
0 & \frac{2 \hat{\sigma}^4}{n}
\end{bmatrix} \tag{2.57}
\]

and the estimated outer-product form of the Fisher information matrix is given by

\[
\hat{R} = \begin{bmatrix}
\frac{1}{\hat{\sigma}^2} X'D_2^2 X' & \frac{1}{\hat{\sigma}^2} X' \frac{S_k}{2\hat{\sigma}_n^2} \frac{(n-q)(Kt-1)}{4\hat{\sigma}^4} \\
\frac{1}{\hat{\sigma}^2} X' \frac{S_k}{2\hat{\sigma}_n^2} & (n-q)(Kt-1)
\end{bmatrix}, \tag{2.58}
\]

where \(D_2^2 = \text{diag}(\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_n^2)\) and \(X\) is \((n \times q)\) matrix of regressors or model matrix, \(S_k\) is the estimated residual skewness, \(Kt\) the kurtosis, and \(1\) is a \((n \times 1)\) vector of ones. That is,
Statistical Data Mining and Knowledge Discovery

\[ Sk = \text{Coefficient of skewness} = \frac{\left( \frac{1}{n} \sum_{i=1}^{n} \hat{e}_{i}^3 \right)}{\hat{\sigma}^3} \]  

(2.59)

and

\[ Kt = \text{Coefficient of kurtosis} = \frac{\left( \frac{1}{n} \sum_{i=1}^{n} \hat{e}_{i}^4 \right)}{\hat{\sigma}^4}. \]  

(2.60)

Hence, the “sandwiched covariance” or “robust covariance” estimator is given by

\[
\text{Cov}(\hat{\theta})_{\text{Misspec}} = \begin{bmatrix}
\hat{\sigma}^2 (X'X)^{-1} & 0 \\
0 & \frac{2\hat{\sigma}^4}{n}
\end{bmatrix}
\begin{bmatrix}
\frac{1}{\hat{\sigma}^4} X' D^2 X & X' \frac{\hat{e}^3}{2\hat{\sigma}^3} \\
X' \frac{\hat{e}^3}{2\hat{\sigma}^3} & (n-q)(Kt-1)
\end{bmatrix}
\begin{bmatrix}
\hat{\sigma}^2 (X'X)^{-1} & 0 \\
0 & \frac{2\hat{\sigma}^4}{n}
\end{bmatrix}.
\]

(2.61)

Note that this covariance matrix in (2.61) should impose greater complexity than the inverse Fisher information matrix (IFIM). It also takes into account presence of skewness and kurtosis which is not possible with AIC, and MDL/SBC. For more on model selection under misspecification, see Bozdogan and Magnus (2003).

Another form of ICOMP is defined by

\[
ICOMP(\text{IFIM})_{\text{Misspec}} = n \log(2\pi) + n \log(\hat{\sigma}^2) + n + 2\frac{\text{tr}(\hat{F}^{-1}\hat{R}) + C_1(\hat{F}^{-1})]}{4\hat{\sigma}^4}.
\]

(2.62)

Similarly, Generalized Akaike’s (1973) information criterion (GAIC) is defined by

\[
\text{GAIC} = n \log(2\pi) + n \log(\hat{\sigma}^2) + n + 2\text{tr}(\hat{F}^{-1}\hat{R}).
\]

(2.63)

For this, see, Bozdogan (2000).


\[
\text{CAICFE} = n \log(2\pi) + n \log(\hat{\sigma}^2) + n + k \log(n) + 2 + \frac{1}{2} \log(\hat{F}) + \text{tr}(\hat{F}^{-1}\hat{R}),
\]

(2.64)

which includes Akaike’s approach and CAICF as special cases. For the application and performance of CAICFE, we refer the reader to Irizarry (2001) for model selection in local likelihood estimation.

We note that the term \( tr(\hat{F}^{-1}\hat{R}) \) in (2.64)-(2.65) is important because it provides information on the correctness of the assumed class of potential models as discussed in White (1982). A fundamental assumption underlying classical model selection criteria is that often the true model is considered to lie within a specified class of potential models. In general, this is not always the case, and often the true model may not be within the model set considered. Therefore, our approach guards us against the misspecification of the probability model as we actually fit and evaluate the models. In this sense, this result is very important in practice, which is often ignored.
2.3.3 AIC and AIC-Type Criteria

AIC for the regression model to be used as fitness values in the GA is given by

\[
AIC(\text{Regression}) = n \log(2\pi) + n \log(\hat{\sigma}^2) + n + 2(k+1).
\]

Similarly, Rissanen (1978) and Schwarz (1978) (MDL/SBC) criterion is defined by

\[
\text{MDL/SBC}(\text{Regression}) = n \log(2\pi) + n \log(\hat{\sigma}^2) + n + k \log(n).
\]

We note that ICOMP and ICOMP(IFIM) are much more general than AIC. They incorporate the assumption of dependence and independence of the residuals and help the analyst to consider risks of both under- and overparameterized models. ICOMP and ICOMP(IFIM) relieves the researcher of any need to consider the parameter dimension of a model explicitly (see Bozdogan and Haughton, 1998 for more detailed comparison.)

### 2.4 A GA for the Regression Modeling

Genetic algorithms (GAs) are a part of evolutionary computing. This is a very fast growing area of artificial intelligence (AI). As it is well known, GAs are inspired by Darwin’s theory about evolution. Simply said, solution to a problem solved by GAs is evolved. Genetic Algorithms (GAs) were invented by John Holland and developed by him and his students and colleagues. This led to Holland’s book *Adaption in Natural and Artificial Systems* published in 1975. In 1992 John Koza used genetic algorithm to evolve programs to perform certain tasks. He called his method “genetic programming” (GP). LISP programs were used, because programs in this language can be expressed in the form of a “parse tree,” which is the object the GA works on.

GA is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope that the new population will be better than the old one. Solutions that are selected to form new solutions (offsprings) are selected according to their fitness value. The more suitable they are the more chances they have to reproduce.

Our implementation of the GA for the problem of model selection in multiple linear regression basically follows Goldberg (1989). Recall that the general regression model can be represented as:

\[
y = X\beta + \varepsilon.
\]  

A GA for the problem of model selection in subset regression models can be implemented using the following steps. For a comprehensive background of GAs and
related topics, we refer readers to Goldberg (1989), Michalewicz (1992), and others. Goldberg’s GA (or called simple genetic algorithm, SGA) contains the following components.

- **A genetic coding scheme for the possible regression models**

  Each regression model is encoded as a string, where each locus in the string is a binary code indicating the presence (1) or absence (0) of a given predictor variable. Every string has the same length, but each contains different binary coding representing different combinations of predictor variables. For example, in a $k=5$ variable regression with a constant term, the string $1 0 1 0 1 1$ represents a model, where constant term is included in the model, variable 1 is excluded from the model, variable 2 is included in the model, and so on.

- **Generating an initial population of models**

  Population size (i.e., number of models fitted) $N$ is an important parameter of GA. Population size says how many chromosomes are in a population (in one generation). If there are too few chromosomes, GA has a few possibilities to perform crossover and only a small part of search space is explored. On the other hand, if there are too many chromosomes, GA slows down. Research shows that after some limit (which depends mainly on encoding and the problem) it is not useful to increase population size, because it does not make solving the problem faster. We first initialize the first population of $N$ random breeding models. Note that the population size $N$, representing the number of models to begin the first run, is chosen by the investigator and not by random. Our algorithm is flexible in that it allows one to choose any population size.

- **A fitness function to evaluate the performance of any model**

  In general we can use any one of the model selection criteria described in Section ?? as the fitness function used in our GA for the regression analysis. However, for the purpose of illustration of the GA and for the brevity in this paper, we restrict our attention to the ICOMP criterion. Analysts can choose any appropriate model selection criterion based on their needs and preferences.

- **A mechanism to select the fitter models**

  This step involves selecting models based on their $ICOMP(IFIM)$ values for composition of the mating pool. After calculating the $ICOMP(IFIM)$ for each of the possible subset models in the population, we subtract the criterion value for each model from the highest criterion value in the population. In other words, we calculate

  $$
  
  \Delta ICOMP(i)(IFIM) = ICOMP(IFIM)_{\text{Max}} - ICOMP(IFIM)_{(i)}
  \tag{2.68}
  $$

  for $i = 1, ..., N$, where $N$ is the population size.

  Next, we average these differences; that is, we compute

  $$
  \overline{\Delta ICOMP(IFIM)} = \frac{1}{N} \sum_{i=1}^{N} \Delta ICOMP(i)(IFIM).
  \tag{2.69}
  $$
Then the ratio of each model’s difference value to the mean difference value is calculated; that is, we compute

$$\frac{\Delta IC_{COMP}(i)_{IFIM}}{\Delta IC_{COMP}(IFIM)}.$$  

(2.70)

This ratio is used to determine which models will be included in the mating pool. The chance of a model being mated is proportional to this ratio. In other words, a model with a ratio of two is twice as likely to mate as a model with a ratio of one. The process of selecting mates to produce offspring models continues until the number of offsprings equals the initial population size. This is called the proportional selection or fitting. There is also rank selection or fitting with $ICOMP$. For this see Bearse and Bozdogan (2002).

**A reproductive operation to perform mating of parent models to produce offspring models**

Mating is performed as a crossover process. A model chosen for crossover is controlled by the crossover probability ($p_c$) or the crossover rate. The crossover probability ($p_c$) is often determined by the investigator. A crossover probability of zero simply means that the members of the mating pool are carried over into the next generation and no offsprings are produced. A crossover probability ($p_c = 1$) indicates that mating (crossover) always occurs between any two parent models chosen from the mating pool; thus the next generation will consist only of offspring models (not of any model from the previous generation).

During the crossover process, we randomly pick a position along each pair of parent models (strings) as the crossover point. For any pair of parents, the strings are broken into two pieces at the crossover point and the portions of the two strings to the right of this point are interchanged between the parents to form two offspring strings as shown in Figure 2.1.

In this case each parent has ten loci. A randomly chosen point along the length of each parent model becomes the crossover point where the models are broken and then reattached to another parent to form two new models. We have a choice of several types of crossover operations. Here, we give just three choices which will suffice for all practical purposes.

**Single point crossover** - one crossover point is selected, binary string from beginning of chromosome to the crossover point is copied from one parent, the rest is copied from the second parent:

Parent A  Parent B  Offspring
11001011  + 11011111 =  11001111

**Two point crossover** - two crossover point are selected, binary string from beginning of chromosome to the first crossover point is copied from one parent, the part from the first to the second crossover point is copied from the second parent and the rest is copied from the first parent:

Parent A  Parent B  Offspring
11001011  + 11011111 =  11011111
An example of the mating process, by means of crossing-over, for a given pair of models (strings).

Uniform crossover - bits are randomly copied from the first or from the second parent:

\[
\begin{array}{c}
\text{Parent A} \quad 0110101100 \\
\text{Parent B} \quad 1101001111 \\
\text{Offspring A} \quad 0110101100 \\
\text{Offspring B} \quad 1101001111
\end{array}
\]

In our algorithm, the user has the option of choosing any one of the above crossover options. Also, the user has the option of choosing what is called elitism rule. This means that at least one best solution is copied without changes to a new population, so the best solution found can survive to the end of the run.

- A random effect of mutation to change the composition of new offspring models

Mutation of models is used in a GA as another means of creating new combinations of variables so that the searching process can jump to another area of the fitness landscape instead of searching in a limited area. We permit mutation by specifying a mutation rate or probability at which a randomly selected locus can change from 0 to 1 or 1 to 0. Thus, a randomly selected predictor variable is either added to or removed from the model.
Depending on the particular crossover and mutation rates, the second generation will be composed entirely of offspring models or of a mixture of offspring and parent models. Models in the second generation then go to produce the third generation; the process continues one generation after another for a specified number of generations controlled by the analyst.

In summary, the outline of the steps of GA is as follows:

1. [Start] Generate random population of \( N \) chromosomes (suitable solutions for the problem).

2. [Fitness] Evaluate the fitness of each chromosome in the population using one of the model selection criteria.

3. [New population] Create a new population by repeating following steps until the new population is complete.
   
   (a) [Selection] Select two parent chromosomes from a population according to their fitness (e.g., ICOMP value)(the better fitness, the bigger chance to be selected).
   
   (b) [Crossover] With a crossover probability cross over the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents. There are three choices of crossover.
   
   (c) [Mutation] With a mutation probability mutate new offspring at each locus (position in chromosome).
   
   (d) [Accepting] Place new offspring in a new population.

4. [Replace] Use new generated population for a further run of algorithm and look for the minimum of the model selection criteria used.

5. [Test] If the final condition is satisfied based on the model selection criteria, stop, and return the best solution in current population.


In the next section, all our computations are carried out using the newly developed graphical user interface (GUI) software in Matlab for the GA subset selection of best predictors. GUI for solving the subset selection model problem is easy to use and very user friendly. The following is the summary of inputs and outputs and usage of GUI.
**List of inputs and usage:**

<table>
<thead>
<tr>
<th>Input</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of generations:</td>
<td>Key in an integer value</td>
</tr>
<tr>
<td>Population size:</td>
<td>Key in an integer value</td>
</tr>
<tr>
<td>Fitness Values:</td>
<td>Check an option in the block:</td>
</tr>
<tr>
<td></td>
<td><em>AIC ICOMP ICOMP(IFIM)</em></td>
</tr>
<tr>
<td>Probability of Crossover:</td>
<td>Key in a real number value from 0 to 1.</td>
</tr>
<tr>
<td></td>
<td>Single Point Two-point Uniform</td>
</tr>
<tr>
<td>Probability of Mutation:</td>
<td>Key in a real number value from 0 to 1.</td>
</tr>
<tr>
<td>Elitism Rule</td>
<td>Check/uncheck the option in the checkbox</td>
</tr>
<tr>
<td>Input Data files:</td>
<td>Button Y: for the response variable Y</td>
</tr>
<tr>
<td></td>
<td>Button X: for the set of predictor variables X</td>
</tr>
<tr>
<td>Go:</td>
<td>To solve the problem</td>
</tr>
<tr>
<td>Reset:</td>
<td>Reset all keyed-in inputs and outputs</td>
</tr>
<tr>
<td>Exit:</td>
<td>Exit the program</td>
</tr>
</tbody>
</table>

**List of outputs and usage:**

<table>
<thead>
<tr>
<th>Output</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. View 2-D/View 3-D buttons:</td>
<td>Shows the 2D/3D plots of criterion values versus generations.</td>
</tr>
<tr>
<td>2. Save Figure:</td>
<td>Pop up the current figure showing in the GUI plot window, then the user can save the pop up figure.</td>
</tr>
<tr>
<td>3. Output in Matlab</td>
<td>Shows the table of generations of GA, Command Window: Shows the fitted chromosomes (Variables) and Binary Strings, and Criterion Score Values.</td>
</tr>
<tr>
<td>4. Output File:</td>
<td>Same as the output in Matlab command window.</td>
</tr>
<tr>
<td>5. Save Figures:</td>
<td>Simply click on view 2-D/View 3-D after the run.</td>
</tr>
</tbody>
</table>

### 2.5 Numerical Examples

We consider two examples to illustrate the implementation of the GA in subset selection of best predictors.

#### 2.5.1 Subset Selection of Best Predictors in Multiple Regression: A Simulation Example

In this example, we generated the values for $y$ and $x_1, x_2, \ldots, x_{10}$, using the following simulation protocol.
We simulate the first five predictors using the following:

\[ x_1 = 10 + \varepsilon_1, \]
\[ x_2 = 10 + 0.3\varepsilon_1 + \alpha\varepsilon_2, \text{ where } \alpha = \sqrt{1 - 0.3^2} = \sqrt{0.91} = 0.9539 \]
\[ x_3 = 10 + 0.3\varepsilon_1 + 0.5604\alpha\varepsilon_2 + 0.8282\alpha\varepsilon_3, \]
\[ x_4 = -8 + x_1 + 0.5x_2 + 0.3x_3 + 0.5\varepsilon_4, \]
\[ x_5 = -5 + 0.5x_1 + x_2 + 0.5\varepsilon_5, \]

where \( \varepsilon_i \) is independent and identically distributed (i.i.d.) according to \( N(0, \sigma^2 = 1) \), for \( i = 1, \ldots, n \) observations, and also \( \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5 \sim N(0, \sigma^2 = 1) \). The parameter \( \alpha \) controls the degree of collinearity in the predictors. Then, we generate the response variable \( y \) from:

\[ y_i = -8 + x_1 + 0.5x_2 + 0.3x_3 + 0.5\varepsilon_i, \text{ for } i = 1, \ldots, n = 100 \text{ observations.} \]

Further, we generate five redundant variables: \( x_6, \ldots, x_{10} \) using the uniform random numbers given by

\[ x_6 = 6 \times \text{rand}(0,1), \ldots, x_{10} = 10 \times \text{rand}(0,1) \]

and fit a multiple regression model of \( y \) on \( X = [x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}] \) for \( n = 100 \) observations, where \( x_0 = 1 \) constant column of \((n \times 1)\) vector of ones.

We expect that the GA run should pick the subset \( \{x_0, x_1, x_2, x_3\} \) to be the best subset selected using the minimum ICOMP value.

The following output from Matlab shows that the GA with ICOMP(IFIM) as the fitness function can detect the relationship and pick the predictors \( \{x_0, x_1, x_2, x_3\} \) to be the best subset chosen as in most of the generations of the GA.

<table>
<thead>
<tr>
<th>Simulation of Collinear Data ( n = 100 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of generations</td>
</tr>
<tr>
<td>Population Size</td>
</tr>
<tr>
<td>Fitness Value</td>
</tr>
<tr>
<td>Probability of crossover</td>
</tr>
<tr>
<td>Elitism (Two point cross over is used.)</td>
</tr>
<tr>
<td>Probability of mutation</td>
</tr>
</tbody>
</table>

**TABLE 2.1**
Parameters of the GA for the simulated model.

In Table 2.1, we show the results of just one run of the GA with the parameters given above.
TABLE 2.2
The results from one run of the GA for the simulated model.

Looking at Table 2.2, we note that GA picks the best subset \{x_0, x_1, x_2, x_3\} very quickly in the second generation, again in the fourth and in fifth generations. Indeed, we also note that in each of the generations the subset \{x_0, x_1, x_2, x_3\} shows up along with one or two redundant variables.

We can repeat this experiment by simulating new \(X - y\) data sets and run the GA many times in order to further illustrate the accuracy and efficiency of the GA.

2.5.2 Subset Selection of Best Predictors in Multiple Regression: A Real Example

In this example we determine the best subset of predictors of \(y = \text{Percent body fat}\) from Siri’s (1956) equation, using \(k = 13\) predictors, \(x_1 = \text{Age (years)}, x_2 = \text{Weight (lbs)}, x_3 = \text{Height (inches)}, x_4 = \text{Neck circumference (cm)}, x_5 = \text{Chest circumference (cm)}, x_6 = \text{Abdomen 2 circumference (cm)}, x_7 = \text{Hip circumference (cm)}, x_8 = \text{Thigh circumference (cm)}, x_9 = \text{Knee circumference (cm)}, x_{10} = \text{Ankle circumference (cm)}, x_{11} = \text{Biceps (extended) circumference (cm)}, x_{12} = \text{Forearm circumference (cm)}, x_{13} = \text{Wrist circumference (cm)}\) using the GA with ICOMP as the fitness function.

The data contains the estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for \(n = 252\) men. This is a good example to illustrate the versatility and utility of our approach using multiple regression analysis with GA. This data set is maintained by Dr. Roger W. Johnson of the Department of Mathematics & Computer Science at South Dakota School of Mines & Technology (email address: rwjohnso@silver.sdsmt.edu, and web address: http://silver.sdsmt.edu/~rwjohnso).

Accurate measurement of body fat is inconvenient/costly and it is desirable to have easy methods of estimating body fat that are not inconvenient/costly. A variety
of popular health books suggest that the readers assess their health, at least in part, by estimating their percentage of body fat. In Bailey (1994), for instance, the reader can estimate body fat from tables using their age and various skin-fold measurements obtained by using a caliper. Other texts give predictive equations for body fat using body circumference measurements (e.g., abdominal circumference) and/or skin-fold measurements. See, for instance, Behnke and Wilmore (1974, pp. 66-67); Wilmore (1976, p. 247); or Katch and McArdle (1977, pp. 120-132). Percentage of body fat for an individual can be estimated once body density has been determined. One (e.g. Siri (1956)) assumes that the body consists of two components: lean body tissue and fat tissue. Letting

\[
D = \text{Body Density (gm/cm}^3\text{)}
\]

\[
A = \text{proportion of lean body tissue}
\]

\[
B = \text{proportion of fat tissue (A} + \text{B} = 1\text{)}
\]

\[
a = \text{density of lean body tissue (gm/cm}^3\text{)}
\]

\[
b = \text{density of fat tissue (gm/cm}^3\text{)}
\]

we have

\[
D = 1/[(A/a) + (B/b)].
\]

Solving for \(B\) we find
FIGURE 2.3
3D Plot: Model Landscape of all Models Evaluated by ICOMP.

\[ B = \left( \frac{1}{D} \right) \left( \frac{ab}{(a-b)} \right) - \left( \frac{b}{(a-b)} \right). \]

Using the estimates \( a = 1.10 \, \text{gm/cm}^3 \) and \( b = 0.90 \, \text{gm/cm}^3 \) (see Katch and McArdle (1977), p. 111 or Wilmore (1976), p. 123) we come up with “Siri’s equation”:

\[
\text{Percentage of Body Fat (i.e., 100} \times B) = \frac{495}{D} - 450.
\]

Volume, and hence body density, can be accurately measured a variety of ways. The technique of underwater weighing “computes body volume as the difference between body weight measured in air and weight measured during water submersion. In other words, body volume is equal to the loss of weight in water with the appropriate temperature correction for the water’s density” (Katch and McArdle (1977), p. 113). Using this technique,

\[
\text{Body Density} = \frac{\text{WA}}{\left( \frac{\text{WA} - \text{WW}}{\text{c.f.}} \right) - \text{LV}}
\]

where \( \text{WA} = \) Weight in air (kg), \( \text{WW} = \) Weight in water (kg), c.f. = Water correction factor (= 1 at 39.2 deg F as one-gram of water occupies exactly one \( \text{cm}^3 \) at this temperature, \( = .997 \) at 76 – 78 deg F), LV = Residual Lung Volume (liters) (Katch and McArdle (1977, p. 115). Other methods of determining body volume are given in Behnke and Wilmore (1974, p. 22).

For this example, we first evaluated all possible subset regression models. Then we picked the top 15 best subset models according to the rankings of the minimum \( \text{ICOMP(IFIM)} \) values. Then, we ran the GA for 100 runs with the parameters given in Table 2.4 and picked the top 10 ranking best subset models according to the minimum value of \( \text{ICOMP(IFIM)} \) to determine if the GA did indeed find the lowest \( \text{ICOMP(IFIM)} \) model in comparison to the all possible subset selection.

The best top fifteen regression models found by the all possible subset selection procedure are given in Table 2.3.
TABLE 2.3
The best models chosen by the lowest fifteen $ICOMP(IFIM)$ values among all possible models for the body fat data.

<table>
<thead>
<tr>
<th>Rank Number</th>
<th>Variables</th>
<th>$ICOMP(IFIM)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1-4-6-7-8-12-13</td>
<td>1473.9065</td>
</tr>
<tr>
<td>2</td>
<td>1-4-6-7-8-9-12-13</td>
<td>1474.5525</td>
</tr>
<tr>
<td>3</td>
<td>1-3-4-6-7-8-12-13</td>
<td>1474.6751</td>
</tr>
<tr>
<td>4</td>
<td>1-4-6-7-8-10-12-13</td>
<td>1475.1721</td>
</tr>
<tr>
<td>5</td>
<td>1-4-6-7-8-11-12-13</td>
<td>1475.2089</td>
</tr>
<tr>
<td>6</td>
<td>1-4-6-7-8-9-10-12-13</td>
<td>1475.5406</td>
</tr>
<tr>
<td>7</td>
<td>1-3-4-6-7-8-9-12-13</td>
<td>1475.6024</td>
</tr>
<tr>
<td>8</td>
<td>1-3-4-6-7-8-10-12-13</td>
<td>1475.7067</td>
</tr>
<tr>
<td>9</td>
<td>1-4-6-7-8-9-11-12-13</td>
<td>1475.8208</td>
</tr>
<tr>
<td>10</td>
<td>1-3-4-6-7--12-13</td>
<td>1475.9539</td>
</tr>
<tr>
<td>11</td>
<td>1-3-4-6-7-8--11-12-13</td>
<td>1476.0138</td>
</tr>
<tr>
<td>12</td>
<td>1-4-5-6-7--12-13</td>
<td>1476.0362</td>
</tr>
<tr>
<td>13</td>
<td>--4-6-7--12-13</td>
<td>1476.116</td>
</tr>
<tr>
<td>14</td>
<td>1-4-6-7-8-10-11-12-13</td>
<td>1476.3913</td>
</tr>
<tr>
<td>15</td>
<td>1-3-4-6-7-8-9-10-12-13</td>
<td>1476.443</td>
</tr>
</tbody>
</table>

TABLE 2.4
Parameters of the $GA$ run for the body fat data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of runs</td>
<td>100</td>
</tr>
<tr>
<td>Number of generations</td>
<td>30</td>
</tr>
<tr>
<td>Population Size</td>
<td>20</td>
</tr>
<tr>
<td>Fitness Value</td>
<td>$ICOMP(IFIM)$</td>
</tr>
<tr>
<td>Probability of crossover</td>
<td>0.5</td>
</tr>
<tr>
<td>Elitism</td>
<td>Yes</td>
</tr>
<tr>
<td>Probability of mutation</td>
<td>0.01</td>
</tr>
</tbody>
</table>

(Uniform crossover is used.)
GA Ranking | Chromosome (Variables) | Binary String | ICOMP(IFIM) |
---|---|---|---|
1 | (1) 1 - - 4 - 6 7 8 - - - 12 13 | 0 1 0 0 1 0 1 1 1 0 0 1 1 | 1473.9 |
2 | (2) 1 - - 4 - 6 7 8 9 - - 12 13 | 0 1 0 0 1 0 1 1 1 1 0 0 1 1 | 1474.6 |
3 | (3) 1 - 3 4 - 6 7 8 - - - 12 13 | 0 1 0 1 1 0 1 1 1 0 0 0 1 1 | 1474.7 |
4 | (4) 1 - - 4 - 6 7 8 - 10 - 12 13 | 0 1 0 0 1 0 1 1 1 0 1 0 1 1 | 1475.2 |
5 | (7) 1 - 3 4 - 6 7 8 9 - - - 12 13 | 0 1 0 1 1 0 1 1 1 0 0 1 1 | 1475.6 |
6 | (8) 1 - 3 4 - 6 7 8 - 10 - 12 13 | 0 1 0 1 1 0 1 1 1 0 1 0 1 1 | 1475.7 |
7 | (9) 1 - - 4 - 6 7 8 9 - 11 12 13 | 0 1 0 0 1 0 1 1 1 1 0 1 1 1 | 1475.8 |
8 | (11) 1 - 3 4 - 6 7 8 - - 11 12 13 | 0 1 0 1 1 0 1 1 1 0 0 1 1 1 | 1476 |
9 | (13) - - - 4 - 6 7 - - - - 12 13 | 0 0 0 0 1 0 1 1 0 0 0 0 1 1 | 1476.1 |
10 | (15) 1 - 3 4 - 6 7 8 9 10 - 12 13 | 0 1 0 1 1 0 1 1 1 1 0 1 1 | 1476.4 |

TABLE 2.5
Top 10 ranking subsets of the best predictors for the body fat data set from 100 runs of the GA.

| RSquare       | =0.741266 |
| RSquare Adj   | =0.733844 |
| Root Mean Square Error | =4.317462 |
| Mean of Response | =19.15079 |
| Observations (or Sum Wgts) | =252 |

TABLE 2.6
Summary of Fit of best subset model.

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimated Coeff.</th>
<th>Std Error</th>
<th>t-Ratio</th>
<th>Prob &gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.63164</td>
<td>6.498054</td>
<td>-0.10</td>
<td>0.9226</td>
</tr>
<tr>
<td>Age</td>
<td>0.0838616</td>
<td>0.029956</td>
<td>2.80</td>
<td>0.0055</td>
</tr>
<tr>
<td>NeckCirc</td>
<td>-0.634546</td>
<td>0.213624</td>
<td>-2.97</td>
<td>0.0033</td>
</tr>
<tr>
<td>Abdo2Circ</td>
<td>0.8808665</td>
<td>0.066639</td>
<td>13.22</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>HipCirc</td>
<td>-0.359215</td>
<td>0.118802</td>
<td>-3.02</td>
<td>0.0028</td>
</tr>
<tr>
<td>TighCirc</td>
<td>0.2826235</td>
<td>0.129812</td>
<td>2.18</td>
<td>0.0304</td>
</tr>
<tr>
<td>ForearmCirc</td>
<td>0.4529919</td>
<td>0.185745</td>
<td>2.44</td>
<td>0.0155</td>
</tr>
<tr>
<td>WristCirc</td>
<td>-1.935856</td>
<td>0.481505</td>
<td>-4.02</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

TABLE 2.7
Parameter estimates of the best subset GA model.
If we had to choose one model from this set, the best subset is the first ranking model according to $ICOMP(\text{IFIM}) = 1473.9$ with the subset \{ $x_1 = \text{Age (years)}$, $x_4 = \text{Neck circumference (cm)}$, $x_6 = \text{Abdomen 2 circumference (cm)}$, $x_7 = \text{Hip circumference (cm)}$, $x_8 = \text{Thigh circumference (cm)}$, $x_{12} = \text{Forearm circumference (cm)}$, $x_{13} = \text{Wrist circumference (cm)}$ \}. Indeed this corresponds to the best subset chosen from all possible subset selection. We further note that GA’s selection corresponds to the top seven best subsets from the results of all possible subsets. This is quite interesting and shows the fact that GA is a highly intelligent statistical model selection device capable of pruning combinatorially large numbers of submodels to obtain optimal or near optimal subset regression models.

Based on our results, the summary of fit and the parameter estimates of the best predictive model are given in Tables 2.6 and 2.7. Figure 2.4 shows the summary of 100 runs of the GA for the body fat data, and Figure 2.5 shows the 3D plot of the model landscape of all models evaluated by the information complexity criterion in 100 runs of the GA for the body fat data.

When we carry out forward stepwise regression analysis on the body fat data set, this approach gives us the full saturated model as the best fitting model, which is not surprising. In other words, stepwise procedure is not able to distinguish the importance of the predictors in the model, since the P-values used in stepwise selection are arbitrary and the F-ratio stopping rule does not have the provision of compensating between the lack of fit and increased model complexity. It does not attempt to find the best model in the model search space.
FIGURE 2.5
3D Plot: Model landscape of all models evaluated by ICOMP(IFIM) in 100 runs of the GA for the body fat data.

Therefore, it is time that researchers start critically thinking of abandoning the use of such procedures, which are less than optimal.

2.6 Conclusion and Discussion
In this paper we have demonstrated that the GA, an algorithm which mimics Darwinian evolution, is a useful optimization tool for statistical modelling. We view each regression model as a chromosome or a genotype within a population of such entities. The value of an information-based model selection criterion is calculated for each model as a measure of the model’s fitness. The collection of all possible fitness values forms a fitness landscape of this “artificial life.” Using the biological concepts of natural selection, crossover, and mutation, the GA searches over this landscape to find the “best” model.

Our GA application to the problem of optimal statistical model selection on the simulated and the body fat indicates that the GA can indeed find the best model without having to evaluate all possible regression models. Compare to all possible subset selection, we evaluated only a small proportion of the total model space in each run.
Question can be asked as to: “Will the GA always be guaranteed to find the best model for all data sets?” In some situations the GA will find “good” models, but may miss finding the overall best one. This is not a failure specific to the GA, but rather a limitation faced by any type of optimization algorithm not using an exhaustive search. Researchers often analyze a large number of variables wherein a very large number of possible models exist. In many data sets, we should bear in mind that a single best model may not exist, but rather a number of equally optimal models are present. A GA is capable, in such a case, of finding at least some of these best models. For example, in the body fat data example, the GA found the best top seven models in the all possible subsets. This is a remarkable achievement.

Another question is: “What are the optimal numbers of both population size and number of generations for use with the GA?” In the literature of GA, there are no clear cut answers to this question. The answer will probably depend on the number of predictor variables and the structure of the data set at hand. Understanding the relationships among these factors requires further investigation (Mahfoud 1994). We are currently investigating these problems by trying different combinations of population sizes and generations, and examining their results carefully to see if they have consistent patterns and results.

The GA approach in combination with information-based criteria is much more likely to uncover the “best” model as well as “better” models than stepwise selection for several reasons. First, our GA approach utilizes an information-based criterion as fitness values to evaluate models instead of treating regression model selection as a problem involving statistical hypothesis-testing, such as P-values used for stepwise selection. Second, the GA approach is not limited to simply adding or removing a single variable at a time. Rather, the GA evaluates models with new combinations of entire sets of variables obtained by evolutionary mechanisms (such as natural selection and crossover) at each generation. Third, the GA is a very flexible optimization algorithm. Several evolutionary mechanisms can be modified by the investigators at different stages of GA. For example, different proportional selection schemes can also be used to determine the combination of genotypes in the population in the next generation (Srinivas and Patnaik, 1994). Finally, and most importantly, a GA is not a “hill-climbing” searching algorithm like forward, backward, and stepwise procedures. A given GA can simultaneously search over multiple areas in the fitness landscape of the solution space.

It is obvious that more research is needed on the application of GAs to statistical model selection, in general, not just for linear regression modelling. We are working on other applications of the GA in other statistical modelling problems such as in logistic and ordinal logistic regression, in vector autoregressive models (Bearse and Bozdogan, 1998), and in multivariate regression (Bearse and Bozdogan, 2002). We encourage researchers to develop GAs for other statistical modelling problems.
Acknowledgments

I was introduced to the genetic algorithms (GAs) by Dr. Hang-Kwang (Hans) Luh in 1996 when he was a post doc in Mathematical Ecology here at the University of Tennessee and worked with me for two years taking my advanced statistical modelling and multivariate courses. I am grateful to Hans for pounding into me the importance of this area in which I was doing statistical modelling with information complexity ICOMP, and marrying it with the GA.

I acknowledge two of my graduate students and thank Vuttichai Chatpattananan and Xinli Bao for creating graphical user interface (GUI) based on my existing Matlab programs of the GA and modifying them. This made the computations much easier.

Finally, I dedicate this paper to Dean Warren Neel for his unending support of my work and area of research since I came to the University of Tennessee in 1990, and for creating a conducive atmosphere to carry out high level research.

References


(VAR) models using the genetic algorithm with informational complexity as the fitness function. *Systems Analysis, Modeling, Simulation (SAMS)*, 31, 61-91.


563-590.


